DOCUMENT RESUME

ED 317 583 TM 014 633

AUTHOR Bezruczko, Nikolaus; Schroeder, David H.
TITLE Artistic Judgment Project I: Internal-Structure Analyses. Technical Report 1989-2.
INSTITUTION Johnson O'Connor Research Foundation, Chicago, IL. Human Engineering Lab.
PUB DATE Dec 89
NOTE 71p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Adults; *Aesthetic Values; Aptitude Tests; *Art Appreciation; Comparative Testing; Evaluation Methods; Factor Analysis; *Factor Structure; *Test Construction; Test Reliability; Test Validity; *Visual Measures
IDENTIFIERS *Artistic Evaluation; Design Judgment Test; Proportion Appraisal; Visual Aesthetic Sensitivity Test; Visual Designs Test

ABSTRACT

An experimental test battery consisting of several tests that measure aspects of artistic judgment was administered to over 1,600 clients of the Johnson O'Connor Research Foundation. The battery consisted of the Visual Aesthetic Sensitivity Test (VAST) of K. O. Gotz (1981); the Design Judgment Test (DJT) of M. Graves (1948); and two tests developed at the Foundation--the Visual Designs Test (VDT), and the Proportion Appraisal. An exploratory factor analysis revealed two major factors for the VDT, and scales named Simplicity and Uniformity were constructed for these factors. The other tests each showed only one major factor. Simplicity, Uniformity, and the DJT showed high reliability. Proportion Appraisal showed substandard reliability, and the VAST showed poor reliability. Sex and age effects were generally small. Intertest correlations were low, indicating that the tests measure largely independent aspects of visual preference and sensitivity. The DJT and the VDT's two scales showed good internal psychometric properties for individual assessment. However, the other tests require further development to meet the Foundation's standards for precision of measurement. Nine figures and 11 tables illustrate the study. A 156-item list of references is included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED317583

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ROBERT KYLE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

ARTISTIC JUDGMENT PROJECT I:

INTERNAL-STRUCTURE ANALYSES

Nikolaus Bezruczko

and

David H. Schroeder

JOHNSON O'CONNOR RESEARCH FOUNDATION

HUMAN ENGINEERING LABORATORY

Technical Report 1989-2

December 1989

TM014633

BEST COPY AVAILABLE

COPYRIGHT © 1989 BY JOHNSON O'CONNOR RESEARCH FOUNDATION, INCORPORATED
ALL RIGHTS RESERVED

Artistic Judgment Project I: Internal-Structure Analyses

Nikolaus Bezruczko and David H. Schroeder

ABSTRACT

An experimental test battery consisting of several tests that measure aspects of artistic judgment was administered to over 1,600 clients of the Johnson O'Connor Research Foundation. The battery contained three tests of visual preference--the Design Judgment Test, the Visual Designs Test, and Proportion Appraisal--and the Visual Aesthetic Sensitivity Test, a test of visual sensitivity to balance and harmony. The Visual Designs Test was developed by Nikolaus Bezruczko at the Foundation, and Proportion Appraisal was developed by previous researchers of the Foundation. The Design Judgment Test and the Visual Aesthetic Sensitivity Test were outside tests.

An exploratory factor analysis revealed two major factors for the Visual Designs Test, and scales named Simplicity and Uniformity were constructed for these factors. The other three tests each showed only one major factor. Simplicity, Uniformity, and the Design Judgment Test showed high reliability. Proportion Appraisal showed substandard reliability, and the Visual Aesthetic Sensitivity Test poor reliability.

Sex and age effects were generally small, although a few substantial findings were noted.

Intertest correlations were low, indicating the tests measure largely independent aspects of visual preference and sensitivity. The direction of the correlations, however, showed surprising consistency with previous factor analytic research.

In conclusion, the Design Judgment Test and the Visual Designs Test's Simplicity and Uniformity scales show good internal psychometric properties for individual assessment. Both Proportion Appraisal and the Visual Aesthetic Sensitivity Test need further development in order to meet Foundation standards for precision of measurement.

CONTENTS

Page

Introduction. 1
Purpose of the Study 1
Historical Perspective 2
Correlates of the Artistic Judgment Tests. . . . 7
The Foundation's Research Perspective on Artistic Judgment. 9

Method. 14
Examinees. 14
Measures 15
Procedures 21
Analyses 21

Results 24
The Artistic Judgment Tests Considered Individually. 24
The Artistic Judgment Battery. 46

Discussion. 48

General Summary and Conclusions 50

References. 52

LIST OF TABLES

		Page
Table 1	Item Statistics for Design Judgment Test	.25
Table 2	Descriptive Statistics for Visual Designs Test . . .	.29
Table 3	Item Statistics for the Simplicity Factor on the Visual Designs Test.	.30
Table 4	Item Statistics for the Uniformity Factor on the Visual Designs Test.	.31
Table 5	Descriptive Statistics for Proportion Appraisal. . .	.34
Table 6	Response Frequencies for Proportion Appraisal. . . .	.35
Table 7	Item Statistics for Proportion Appraisal, Consensus Scoring.	.37
Table 8	Item Statistics for Proportion Appraisal, .67 Scoring.	.38
Table 9	Item Statistics for Visual Aesthetic Sensitivity Test	.43
Table 10	Descriptive Statistics for Three Subscales of the VAST	.44
Table 11	Correlations Among the Artistic Judgment Tests . . .	.47

LIST OF FIGURES

		Page
Figure 1	Sample Item from Design Judgment Test	16
Figure 2	Sample Item from Visual Designs Test.	17
Figure 3	Sample Item from Proportion Appraisal	19
Figure 4	Item Similar to the Items on the Visual Aesthetic Sensitivity Test.	20
Figure 5	Age Curve for Design Judgment Test.	27
Figure 6	Age Curve for Simplicity.	32
Figure 7	Age Curve for Uniformity.	33
Figure 8	Age Curve for Proportion Appraisal, Consensus Scoring	40
Figure 9	Age Curve for Proportion Appraisal, .67 Scoring	41
Figure 10	Age Curve for Visual Aesthetic Sensitivity Test.	46

ACKNOWLEDGMENTS

We gratefully acknowledge the cooperation and assistance of the Washington, DC, New Orleans, Dallas/Fort Worth, Houston, Denver, Seattle, and Philadelphia offices in collecting the data on which this report is based. Without the conscientious attention of the test administrators and directors of these offices, this study would not have been possible.

We compliment the research office staff for their expertise and efficiency in processing these data. Their commitment to accuracy and precision contributed substantially to the validity of our findings.

We extend our sincere appreciation to Professor Hans J. Eysenck of the University of London and to Professor K. O. Götz of the Düsseldorf Academy of Fine Arts. Professor Eysenck provided us with research literature regarding the Visual Aesthetic Sensitivity Test and helped us contact Professor Götz, developer of the test. Professor Götz expressed interest in the study and generously contributed the printed test books that were administered in the testing offices.

INTRODUCTION

Purpose of the Study

In the 1930s, the Johnson O'Connor Research Foundation (JOCRF) began to test persons' preferences for different types of visual designs. The area of study was termed artistic judgment, and the Foundation thought that one or more aptitudes, revealed in visual preferences, underlay satisfaction and success in artistic occupations. The McAdory Art Test was first used to measure visual preferences and was followed later by Proportion Appraisal, a test developed by the Foundation. In 1948, The Psychological Corporation published the Design Judgment Test, which was administered regularly, together with Proportion Appraisal, in the Foundation for many years. In the 1940s, however, questions arose concerning the validity of Proportion Appraisal (see Technical Report 99). Specifically, questions arose whether the test measures differences in the preference for proportions or responses are simply an artifact of the item format. This issue was part of a larger concern whether preference judgments on the test are associated with a fundamental appreciation of beauty. By the 1960s, concerns also arose regarding the unidimensionality and validity of the Design Judgment Test (Eysenck, 1967, 1970; Eysenck & Castle, 1971). In the 1970s, the Foundation dropped first Design Judgment and then Proportion Appraisal from the battery. Similarly, outside the Foundation, commercial artistic judgment tests went out of print. A review of Buros's Mental Measurement Yearbooks indicates that currently there are no validated tests of artistic judgment available commercially (Conoley & Kramer, 1989).

Continued interest in understanding artistic judgment, as well as the need for an effective art aptitude test in the battery, has led the Foundation to initiate further research on artistic judgment. In early 1985, Nikolaus Bezruczko and others at the Foundation established a plan for research on artistic judgment that consisted of the following steps:

1. critically reviewing the research literature on artistic judgment;
2. as indicated by the literature review, developing a new test to measure artistic judgment;
3. administering a battery of the most promising artistic judgment tests in a construct validation study; and
4. administering the same battery of tests to a sample of professional artists in an occupational validation study.

These plans have now been carried out by Bezruczko and others. The review of the literature was published as Technical Report 1988-1. The construction of a new test, the Visual Designs Test, was described in part in Statistical Bulletin 1986-5 and the appendix to Technical Report 1988-1, in addition

to the Method section of this report and Research Memorandum 1989-1. The present report covers the internal-structure analyses (reliability, item analysis, and so on) of four artistic judgment tests that were administered in the construct validation study. In later reports we will address the construct validity of the four tests and the results of an occupational validation study with professional artists.

Before we present the internal-structure findings, we first provide a historical context for artistic judgment testing by briefly reviewing the major artistic judgment research themes of the past century. This presentation will include a review of relevant concepts and terms and the rationale for selecting the four tests that we used in this study. The technical features of these tests appear later in this report, in the Measures portion of the Method section.

Historical Perspective

Within modern psychology, research has involved four major approaches to the study of artistic judgment, namely: (a) psychophysical measurement, (b) rationally derived artistic judgment tests, (c) factor-analytic research, and (d) psychobiological investigations. These perspectives and the findings associated with them are briefly described below.

Psychophysical Measurement

Although not widely known to contemporary researchers, a psychophysical approach to artistic judgment was first taken in the nineteenth century by Fechner, in a series of experimental studies investigating preferences for proportions in visual figures (Fechner, 1865, 1876). In these studies, he identified a logarithmic relationship between changes in the proportions of the sides of rectangles and observer preference judgments. Although Fechner's studies did not lead to a great deal of further research, his methods have been widely applied in the study of perception and neuropsychology and form the basis for modern educational and psychological scaling methods. He adapted a classical philosophical principle known as "unity in variety" (Bosanquet, 1917) in his research that continues to influence empirical aesthetic theory. In the 1920s, Thurstone adapted Fechner's psychophysical techniques to ability and attitude testing, noting their applicability to the study of aesthetics (Thurstone, 1927a, 1927b, 1927c, 1931, 1948, 1954).

Artistic Judgment Tests

By the beginning of the twentieth century, research on art ability was focused on two areas: art production and artistic judgment. Work directed toward measuring art production concentrated on tests of drawing ability (Kline & Carey, 1922; Lukens, 1896; Manuel, 1919; Thorndike, 1913), while measurement

of artistic judgment focused on sensitivities and discrimination, skills viewed critical to persons' artistic capabilities. An example of an early, widely used test of artistic judgment is the McAdory Art Test (McAdory, 1929).

The McAdory Art Test consisted of 72 items, each of which presents four hand-rendered illustrations of common everyday objects. The objects were selected from a variety of sources as described in the following quote.

> Magazines supplied designs for automobiles, furniture, clothes, and rugs. Department stores were a source for items relating to dress accessories and clothes of men, and children. From art museums came designs in ceramics, silver, furniture, metal work, textiles, lace, printing and book-cover designs, sculpture, and furniture. (McAdory, 1929, p. 2)

In each item, one of the four representations is artistically the best. McAdory (1929) reported that she manipulated three elements in constructing the items: "shapes and line arrangements, values of dark and light, and color-use of hue, value, and chroma" (p. 5). Her criterion for scoring the test was the consensus judgment of 100 professional artists.

The test was reliable ($r_{tt} > .90$), but the scores correlated with art training and experience (McAdory, 1929; also see Carroll, 1933), suggesting that the test is sensitive to skill acquisition and thus not truly an aptitude. The Foundation included the McAdory Art Test in the battery for several years, but in 1940 O'Connor noted that the content of the items had become old-fashioned, and its use was discontinued. Around the same time, Meier and Seashore (1930) introduced the Meier-Seashore Art Judgment Test, similar to the McAdory Art Test except that it relied exclusively on the use of authentic art reproductions. Meier (1942c) later revised the test, which was then published by The Psychological Corporation as the Meier Art Test.

Meier, like McAdory, emphasized an artistic judgment construct based on the preference for illustrations that differed in their spatial composition (Meier, 1928). In his test, however, Meier matched pairs of reproductions of classical art works rather than the illustrations of common objects as in the McAdory Art Test. He manipulated one of the reproductions in each pair so that it contained design flaws. According to Meier, the design flaws generally involved art principles relating to harmony, balance, or rhythm: the basic elements in his concept of composition. Test scores were moderately reliable (.70), although they correlated with age and amount of previous training, suggesting that like the McAdory Test, the Meier test was influenced by experience (Barrett, 1949; Crannel, 1953).

In 1948, The Psychological Corporation published an artistic judgment test that did not rely on authentic or historical art reproductions or objects encountered in everyday experience, diminishing the influence of experience. This test, named the Design Judgment Test, used nonrepresentational designs purportedly constructed on the basis of several aesthetic principles (Graves, 1948). The reliability of the test was high (> .90), and the test was widely used. Initial studies by Graves supported its validity, although researchers (Child, 1964; Eysenck, 1970; Götz & Götz, 1974) rejected its use on psychometric and validity grounds.

Another area of artistic judgment research concerns a test by Welsh (1949). In the course of developing a figure preference test for personality assessment, he and Barron (Barron & Welsh, 1952) discovered that artists tended to have distinctive preferences for certain figures on the test. Subsequent research indicated that the same preferences also correlated somewhat with general creativity (Barron, 1965; Welsh, 1975). In their analysis of test responses, Barron and Welsh (1952) attributed the differences between artist and nonartist preferences to the complexity of the figures, with artists preferring the more complex ones. Based on our assessment of the artistic judgment literature, however, we find that their interpretation needs revision.

Brighouse (1939b), for example, found in a study of visual perception that artists preferred simpler rather than more-complex paintings. In a study of the preference for polygons, Harsh, Beebe-Center, and Beebe-Center (1939) found a bipolar simplicity-complexity factor that was replicated by Eysenck (1941b). Eysenck and Castle (1970) found that artists tend to score toward the simplicity pole of this factor. Their finding is further supported by subsequent research (Fritzky, 1963; Vanderplas & Garvin, 1959). Eysenck (1981) discussed the possibility that the preference for simplicity is related to a K factor he found in earlier research (Eysenck, 1941b).

In view of this research, we believe the discriminating factor in the Barron-Welsh Art Scale is probably something other than the preference for complexity. Since Barron and Welsh's interpretation of complexity was not based on a controlled manipulation of the designs, we offer an alternative explanation, namely, that the artist-nonartist differences on the Barron-Welsh Art Scale are due to differences in the degree of emotionality and abstraction between keyed and nonkeyed designs. This explanation is based on research by Goude (1972) indicating that artists prefer designs with higher emotional content than nonartists (see also Pickford, 1948) and research by Getzels (1976), who found that artists prefer abstract images (see also Child, 1965). The preference for abstractness is consistent with research by Peel (1944) and Francés and Voillaume (1964). A visual inspection of the Barron-Welsh Art Scale indicates that

many of the keyed designs are abstract and have more emotional content than alternative designs.

Another empirical perspective on artistic judgment testing involves the preference for proportion and is derived from research in the eighth century B.C., when Pythagoras discovered the importance of ratio, or proportion, for music. For centuries art theorists have tried to establish that the liking for visual designs is influenced by a principle of spatial organization called the golden section. With reference to a design or figure, the golden section is said to occur when the ratio of the width (W) to the length (L) is .618. This value has the unique mathematical property that W/L = L/(W+L) and is also related to a mathematical structure known as the Fibonacci sequence (see Berlyne, 1971). With regard to art, the golden section refers to a hypothesis that the more a design or figure conforms to the golden section proportion, the better it will be liked.

Despite widespread belief in the golden section's importance for art, empirical support for the hypothesis has been inconsistent. Thorndike (1916) found support for it in his test of art appreciation, although Berlyne (1971) speculated that "the golden section may actually not be any more important than many neighboring proportions" (p. 229). Other research (Eysenck & Turnstall, 1968) also discounts its influence on artistic judgment.

Independent of the golden-section hypothesis, the influence of proportion on artistic judgments led the Foundation to construct a test called Proportion Appraisal. Developed in the 1930s, the test was designed to remedy several problems with the McAdory Art Test, such as the length of testing time and the lack of a unidimensional construct underlying the scale.

Proportion Appraisal consists of sets of geometric figures that vary only in their proportions (e.g., sets of rectangles of varying proportions, ellipses of varying proportions, and so on). Proportion Appraisal was administered regularly for many years in the Foundation's testing program, although certain questions regarding its validity as an artistic judgment test and the interpretation of scores on the test were never resolved.

Factor-Analytic Research

Simultaneous with the development of artistic judgment tests were factor analytic studies by Eysenck (1940a, 1941a, 1957, 1968, 1971). His early studies of preference judgments indicated a strong general artistic judgment factor that extended across a wide range of visual stimuli and perceptual modalities and was consistent with O'Connor's position that "some general niceness of sense perception" (O'Connor, 1948, p. 5) underlies interest in fine art. O'Connor also found evidence for this general aptitude across modalities and aesthetic activities including musical

performance, amateur and commercial photography, and jewelry construction (O'Connor, 1948). The claim of a general aptitude, however, is not generally supported by other research. Although Granger's (1955) research supported Eysenck's finding of a general artistic judgment factor, Child (1964) failed to find the factor. Within the Foundation, Hill (1968) found a correlation of only .14 between pitch discrimination and color discrimination, two attributes that O'Connor believed to be manifestations of this ability. More recently, researchers have developed an artistic judgment test (Götz, Lynn, Borisy, & Eysenck, 1979) based on a concept of good design that is derived from Eysenck's theory of a general artistic judgment aptitude. This is the Visual Aesthetic Sensitivity Test, which has been applied successfully with adults and children (Eysenck, 1983).

Several findings from Eysenck's research should be noted. First, he reports finding a general consensus among both artists and nonartists regarding which designs are more and less liked, which provides a basis for assigning objective values to art works. Second, Eysenck found a general individual-difference factor that he called "T" (or "Taste") that represents the extent to which persons agree with the consensual values. Third, he found another factor called "K" that distinguishes between the preference for bright, colorful art, characteristic of contemporary works, as opposed to the darker, gloomier style of classical artworks which has significant personality correlations. Eysenck also found that the K factor, when measured with a test contrasting simple and complex polygons, shows artists to prefer simpler designs (Eysenck & Castle, 1970), a finding supported by other research (Brighouse, 1939b; Eysenck, 1972a; see earlier discussion of the Barron-Welsh Art Scale).

In 1985, when we reviewed our perspective on artistic judgment, Eysenck's findings with his K factor encouraged us to explore the complexity of designs as a means of distinguishing between artists and nonartists. Later in this report, we describe the Visual Designs Test, a test derived from the K factor for this purpose.

Psychobiological Approach

By the end of the 1960s, artistic judgment test development had come to a standstill. The last major test, the Barron-Welsh Art Scale, was published in 1963 (Welsh & Barron, 1963), and today all the major tests (i.e., the McAdory, Meier, Design Judgment, and Barron-Welsh tests) are out of print. During the intervening years, however, a Canadian researcher named Berlyne (1971, 1974) took a different approach to the study of artistic judgment. Berlyne and his colleagues were psychobiologists who did not have a particular interest in test development or psychometrics. Rather, their main emphasis was on understanding artistic judgment as a functional, adaptive aspect of human behavior. This point of view led them to conduct an extensive

series of studies investigating persons' preferences for visual patterns. Among their findings, one that relates to Eysenck's earlier factor analytic studies is the existence of a curvilinear relationship between the complexity of a design and viewer liking. Although Eysenck found nonartists to prefer more-complex designs, Berlyne discovered that increases in the complexity of designs increase liking for designs only up to a point. Beyond that point, further increases in complexity cause decreased liking.

The implications of a curvilinear relationship between preference and complexity for distinguishing between artists and nonartists on an artistic judgment test has not been fully explored. We speculate that the curves for artists and nonartists have different peaks and that this difference can be identified in responses to test items. More specifically, Berlyne's research may have implications for construction of artistic judgment test items.

Correlates of the Artistic Judgment Tests

Relationships Among the Tests

Millman and Chang (1966) found positive correlations between the Design Judgment Test and, respectively, the Barron-Welsh Art Scale (.46) and the Meier Art Test (.19). Carroll and Eurich (1932) reported a correlation between the McAdory Art Test and the Meier-Seashore Art Test of .27 (.34 after correcting for attenuation), which was replicated by Wallis (Siceloff & Woodyard, 1933). These results are consistent with between-test correlations reported by others ranging from .21 to .51 (Anderson 1951; Barrett 1949; Bolton 1955; Crannell 1953; Dewar, 1938; Eysenck, 1972a), although Child (1964) found a correlation of only .12 between an artistic judgment test by Bulley (1951; adapted by Child so that it consisted of a filmstrip with 19 pairs of art objects) and Design Judgment. In view of the low intertest correlations, Eysenck offered the possibility that "the effects of specialized teaching, rather than aesthetic sensitivity" (Eysenck, 1972a, p. 556) may determine performance on these tests. In other words, these tests appear "more in the nature of achievement tests than of native ability tests, i.e. scores are influenced by the nature and amount of previous training" (Barrett, 1949, p. 400). Child (1964) supports this opinion with evidence that the full (90-item) Design Judgment Test is positively correlated with freshman grade-point average (.16), an English achievement test (.16), and the Test of Developed Ability (see Coffman, 1957; Dyer, 1957), including subtests of science knowledge (.14) and science abilities (.22).

Relationships With Other Variables

Cognitive ability variables. Although the validity of artistic judgment tests has not been fully determined, a

plausible pattern of relationships for many of them with other variables is evident. Child (1964) found that the Bulley Test (Bulley, 1951) correlated .36 with the Spatial Relations Test of the Yale Aptitude Battery (Crawford & Burnham, 1946). Likert and Quasha (1948) and Woods (1953) both found a correlation between success in art and spatial aptitude as measured by the Minnesota Paper Form Board Test. Schroth (1967), however, was unable to replicate their finding using the Differential Aptitudes Test (DAT) spatial subtest. He speculated that the DAT measures a more complex aspect of spatial relations than is needed for success in art. Winner (1988), however, reported that art and math majors out-performed psychology majors on a mental paper folding task associated with spatial ability, and Rosenblatt and Winner (1988) found artists to have superior visual memory.

In Eysenck's 1972 study, artists scored slightly higher than nonartists on a vocabulary test, which is consistent with other research indicating a modest positive correlation between artistic judgment and general intelligence measures. Bolton (1955) found the Design Judgment Test to correlate between .10 and .19 with intelligence. Child (1964) reported a positive correlation between Design Judgment and the verbal scale of the Scholastic Aptitude Test (.18), and Meier (1942b) reported correlations between the Meier Art Test and intelligence ranging from -.14 to .28. Terman and Oden (1947) found a significant relationship between general intelligence as measured by the Stanford revision of the Binet Test and the creative achievement of great artists.

Personality variables. Using Cattell's Sixteen Personality Factor Test, Csikszentmihalyi and Getzels (1973) found that art students scored high on scales measuring sensitivity and self-sufficiency but low on scales measuring emotional stability and conformity to norms. Barron and Welsh (1952) speculated and Barron (1953) confirmed that scores on the Barron-Welsh Art Scale are associated with two personality characteristics, originality and independence of judgment. Barron (1953) also found art students to be independent and unconventional. Eysenck (1972a) and Eysenck and Castle (1970) found art students significantly more introverted and neurotic than nonart students on the Eysenck Personality Inventory (Eysenck & Eysenck, 1964), which is consistent with Götz and Götz (1979a), who found male artists more introverted and neurotic than male nonartists. According to Eysenck (Eysenck, 1988; see also Götz & Götz, 1979b), artists are also much higher on his "Psychoticism" scale than nonartists, which indicates that artists are more tough-minded than nonartists. According to Eysenck and Eysenck (1976), high scorers on this scale may be characterized as "solitary, troublesome, and aggressive, and they like odd and unusual things" (p. 332). These findings are supported by Götz et al. (1979), who found scores by a nonartist sample on a test of aesthetic judgment to be inversely correlated with psychoticism scores.

Some researchers note a higher incidence of mental disturbance for artists than other persons (Andreasen, 1987; Ellis, 1926; Juda, 1949; McNeil, 1971), which is particularly interesting considering the Neale and Cromwell (1969) finding that schizophrenics prefer randomly generated polygons with lesser amounts of complexity. As noted earlier, Eysenck and Castle (1970; see also Eysenck, 1972a) found that artists preferred polygons with lesser complexity.

The Foundation's Research Perspective on Artistic Judgment

These studies provide a substantial foundation for the study of artistic judgment. Fechner's psychophysical studies imply a mathematical structure to visual preference, while the studies of artistic judgment tests, including the Foundation's Proportion Appraisal, have identified several apparently distinct dimensions that discriminate between artists and nonartists. Eysenck's research articulates an underlying factor structure, a perspective that is reinforced by Berlyne's studies. On the other hand, a number of unanswered questions remain for this area. Some of these questions are:

1. Are test stimuli culturally specific or general (unbiased) across cultures?
2. What evidence is there that these tests measure aptitude for art?
3. Following Question 2, has anyone shown that these tests measure their intended constructs?
4. Do the correlations with age and art training threaten the validity of the tests?
5. Why do expert judgments of aesthetic value correlate poorly?
6. Do the test stimuli encompass "good" art?
7. Are the artistic stimuli on the tests too complex (multifaceted) to be studied systematically? and
8. When are the differences between personal preferences and artistic judgments important for the study of artistic aptitudes?

In the Foundation study to be reported here, we attempted to address some of these issues. Through the controlled production of visual designs, we constructed a new artistic judgment test designed to be relatively culture-free. We included this test in a battery of artistic judgment measures along with a number of indicators of validity. Specifically, we examined convergent validity with choice of college major, an artistic background questionnaire, and an occupational interest inventory. We examined discriminant validity with age and cognitive ability measures. Importantly, we tested both Foundation examinees and an independent sample of professional artists, to see if the tests discriminate between these two groups.

In this report, we present our findings regarding the internal structure of the four artistic judgment tests that we used. Specifically, we address whether these tests measure one or more traits accurately. Our findings regarding construct and occupational validity will be presented in later reports.

Tests Selected For the Study

Based on a survey of the research literature and a consideration of the problems in past artistic judgment research, we chose four tests for this study. They are the Design Judgment Test, the Visual Designs Test, the Visual Aesthetic Sensitivity Test, and Proportion Appraisal. A brief description of the tests and their function in the study follows. The Method section of this report presents the technical characteristics of the tests, as well as sample test items.

Design Judgment Test (DJT; Graves, 1948). Historically, the Design Judgment Test is the most thoroughly researched and widely used test of artistic judgment. In each item, "one design was organized in accordance with the fundamental principles of art structure" (Graves, 1948, p. 2), which Graves considers to be unity, dominance, variety, balance, continuity, symmetry, proportion, and rhythm, although the specific influence of these principles on the construction of items has not been explicated by the test author. An innovation in the test is the use of nonrepresentational designs to minimize cultural influences on responses. Eysenck (1967) factor analyzed the test, however, and found that, as a whole, the test is not unidimensional and, in Eysenck's judgment, should not be used as a test of artistic judgment. On the other hand, Eysenck (1972b) later found a subset of 22 items from the original 90 that are unidimensional and may be related to artistic judgment. (Henceforth, when we refer to the Design Judgment Test [DJT], we mean the 22-item subset.)

The low correlation of the test with intelligence test scores and art training (Bolton, 1955) and the tendency for artists to score higher than nonartists (Graves, 1948) suggest that the DJT may measure a distinct aptitude for artistic work. Questions remain, however, regarding the generality of the underlying construct across diverse groups of artists and the association of the DJT with other tests. The reliability of the 22-item subset also needs to be established. These considerations led to the inclusion of the DJT in our artistic judgment study.

Visual Designs Test (VDT; Worksample 736 C). The VDT is an experimental test developed by Nikolaus Bezruczko at the Foundation to address the problems of culture-boundness, operational definition, and aesthetic value of items associated with previously developed tests. With the exception of Birkhoff's (1932) work, the operational definitions for the factors that underlie tests of artistic judgment either were

never explicit or were totally lacking in objectivity, and artists oftentimes commented that the items lacked aesthetic value. The VDT addresses these problems by employing a reasonably culture-free construct, namely, the visual preference for lines oriented in space. In a VDT item, the spatial arrangement of these lines is manipulated by varying two factors that previous research indicates are associated with artistic preference: complexity (or, conversely, simplicity) and order (sometimes referred to as redundancy) in a manner patterned after a contemporary art style.

Complexity has a long history in aesthetic theory, first appearing in empirical research in Fechner's studies of the preference for the golden section and reappearing in Birkhoff's research in which artistic value was hypothesized as a function of the ratio of order to complexity (see discussion in Technical Report 1988-1). Contemporary researchers widely agree that complexity is related to artistic judgment, although they do not agree on the nature of the relation. Researchers have proposed several methods for manipulating complexity in test items, although none have received wide acceptance. In this study we use in a modified way Birkhoff's original operational definition of complexity, which is described in the Method section.

Attneave (1954, 1959) identified redundancy, or repetition among the parts of a design, as an influence on the design's perception as "good" (good as interpreted by Gestalt psychologists). He (Attneave, 1959) demonstrated that repetition produces perceived order, although he did not investigate its influence on preference empirically. His perspective, however, is supported by empirical research (Garner, 1970; Hochberg & McAlister, 1953), which indicates that figural "goodness" is in part a function of redundant information in a stimulus.

With this perspective on complexity and order, Bezruczko constructed a testing device in which the visual information regarding complexity and order in a design is conveyed to a viewer during a preference judgment. Because artists and nonartists are expected to differ in their evaluations of complexity and order, test items constructed on this principle should distinguish between them.

Regarding the aesthetic value of the items in previous artistic judgment tests, Eysenck (1988) has frequently noted the negative reactions of artists to tests that use polygons, random shapes, and the like. Artists comment that the designs are poorly done and generally lack aesthetic value. To avoid this problem with the VDT items, Bezruczko constructed items using a minimalist art style called neoplasticism. This style was introduced by Mondrian and Van der Leck earlier in this century (Lemoine, 1987; see also Champa, 1985) and is characterized by simple relationships between line and space, presumably capturing an underlying essence of visual art experience. Within this

style, then, Bezruczko systematically manipulated the design features of complexity and order while maintaining a degree of overall aesthetic value.

The result is that the VDT has designs that vary in complexity and order in a manner that we believe is relatively independent of culture and reasonably pleasing to artists. The specific item-construction procedures are presented in the Method section.

The VDT designs were pilot-tested in an earlier study (Statistical Bulletin 1986-5; Test Information Bulletin 1986-10). In the present study, we evaluated the internal structure (reliability, item-total correlations, and so on) of the test, as described in the Results section of this report. We also examined the construct and occupational validity of the test, to be covered in later reports.

While the primary purpose of this analysis of the VDT is to identify its psychometric properties, the clear fashion with which complexity is manipulated across varying levels of order in the test suggests the results could clarify an area of confusion in artistic judgment research. Some studies find artists preferring more-complex designs (Barron & Welsh, 1952; Biaggio & Supplee, 1983; Eisenman, 1966; Eisenman & Rappaport, 1967; Munsinger & Kessen, 1964; Osborne & Farley, 1970; Taylor, & Eisenman, 1964; Vitz, 1966), while other studies indicate that artists prefer less-complex designs (Birkhoff, 1932; Brighouse, 1939b; Eysenck, 1940b, 1941a, 1970, 1972a; Vanderplas & Garvin, 1959). This inconsistency needs to be resolved for the development of art aptitude tests to proceed.

Proportion Appraisal (PA). Originally relying on the McAdory Art Test as its test of artistic judgment, the Foundation became concerned that McAdory's simultaneous manipulation of multiple elements within items might not measure a unitary trait (Technical Report 23; Technical Report 44). Also, O'Connor (1940) noted that, over time, the visual content of the McAdory items (adult and children's clothing, household goods, and so on) had become old-fashioned looking and distracting to examinees. On the basis of these issues, along with the belief that differences in the preference for proportion represented a fundamental aspect of artistic judgment, the Foundation constructed a new test. Sets of simple geometric designs were constructed that varied only in their proportions (width/length) from 1/4 to 1/2 to 2/3 to 3/4. (The 1/4 designs were rarely chosen and were later deleted from the test.) For each set, examinees choose the design that they prefer. Scores are based on whether one prefers the moderate 2/3 proportion or the more-elongated 1/2 proportion.

By including PA in this study, we hope to ascertain whether the original scoring method based on consensus of opinion

(utilizing the ratio .67, although sometimes .75 and .50) is still reliable and determine whether the test can effectively discriminate between nonartists and professional artists.

Visual Aesthetic Sensitivity Test (VAST; Götz, 1981). This is a test developed by Götz and Eysenck and their colleagues (Götz et al., 1979) based on their assumption that good art shares the common characteristics of balance and harmony. According to their perspective, all elements in an aesthetic design should fit together in producing a pleasing image. Therefore, the ability to recognize balance and harmony in designs is a prerequisite for artistic pr duction, and the VAST is designed to measure this ability. Although this rationale is similar to the underlying basis for the Meier Art Test and the McAdory Art Test, the VAST is distinguished by its use of pairs of nonrepresentational designs that have been evaluated by a group of eight professional artists. Only items that the artists agreed upon unanimously are included in the test.

Within each pair of designs in the test, one corresponds better to Eysenck's concept of "good gestalt," (Eysenck, 1942) based on his interpretation of principles from Gestalt psychology (Koffka, 1935; Köhler, 1920). These principles emphasize the symmetry, simplicity, and regularity (among other characteristics) of "good" designs and are supported by Eysenck's extensive factor analytic research (Eysenck 1940a, 1941a), in which he found a general factor (named "T") that occurs across a wide range of preferences and modalities. The VAST is designed to measure T, and some research indicates it is psychometrically sound. In an investigation of its cultural independence with Japanese and English children and adults, Iwawaki, Eysenck, and Götz (1979) found VAST scores for Japanese children and adults superior to an English sample, suggesting that the answer key is valid in non-Western cultures. In a similar study, Chan, Eysenck, and Götz (1980) found no evidence of greater similarity between Chinese and Japanese groups, in comparison to a group of English children, further supporting the cultural independence of the test construct.

Although the rationale for the VAST is clear and plausible, it has yet to be validated authoritatively. The verification of the key by professional artists was encouraging, although the group was small, and the nonartist studies reported to date use relatively restricted populations (e.g., college students, children). Although this test is promising, the Foundation needs additional data regarding reliability and validity with a larger artist sample.

Expected Findings

Previous research has indicated that different artistic judgment tests appear to measure different variables. Eysenck's research, however, suggests that one should find one factor (his

T factor) on which the preferences of artists and nonartists tend to agree and possibly a second factor (K) that distinguishes between artists and nonartists (artists prefer simplicity). In previous research, Eysenck (1972a) found that the DJT does not correlate with a polygon test of T, although it correlates negatively with a polygon K factor, indicating that the preference for simplicity is inversely related to the preference for asymmetry (the keyed response on the DJT). Thus, the DJT should correlate negatively with the VDT, which is related to K. (Eysenck's correlation between DJT and K, however, is inconsistent with his research indicating that artists prefer asymmetry, which suggests that the DJT should correlate positively with K.)

One of the tests in this study, PA, is based on an idea of general consensus of agreement and, thus, is similar in concept to the VAST, which measures a general preference factor (Eysenck's T). Therefore, PA and the VAST should correlate positively. Because DJT does not correlate positively with T (Eysenck, 1972a), we do not expect PA or VAST to correlate positively with it.

Artists are expected to score higher than nonartists on the DJT (Graves, 1948), as well as on the VDT (in the direction of preference for simplicity). Because the DJT is negatively correlated with PA (Statistical Bulletin 1975-6; Technical Report 792), artists should score low on PA.

METHOD

Examinees

The examinees in this study were paying clients of the Foundation's aptitude-assessment service. The examinees took a battery of experimental artistic judgment tests for this study in addition to the standard JOCRF aptitude battery.

A total of 1,686 examinees completed one or more of the tests in the artistic judgment battery. Of these examinees, 53.4% were male and 46.6% were female. The age of the examinees ranged from 14 to 54 years with an average of 26 years (SD = 10.04). The median age was 22 years, indicating a moderate skewness toward older examinees. Racially and economically, examinees were generally white and middle to upper-middle income. Most of the examinees were college-educated or college-bound.

Geographically, 482 of the examinees were tested in Foundation offices in the eastern United States (28.6%), 870 examinees were tested in southern Foundation offices (51.6%), and 334 of the examinees were tested in western offices (19.8%).

Measures

The tests in the artistic judgment battery were the Design Judgment Test, the Visual Designs Test, Proportion Appraisal, and the Visual Aesthetic Sensitivity Test. Description of the tests, their method of scoring, their technical characteristics, and sample items from the tests follow.

Design Judgment Test (Graves, 1948). On each item in the DJT, examinees are presented with two or three visual designs printed on a page in a two- or three-color format (black on a white background or black and white against a light green background). Examinees are directed to choose the one design from each group that they most prefer, that is, that they like the most. Examinees' scores consist of the number of times their choice corresponds to the choice considered artistically superior.

In 1948, Graves reported a median alpha reliability of .86, based on 14 samples of art and nonart students. In validation studies, he found that artists scored more than 20 points higher than nonartists on the DJT (out of 90 items). As noted earlier, Eysenck (1970; Eysenck & Castle, 1971) failed to replicate the latter finding in other samples, although Eysenck and Castle (1971) found a 13-item subset of the DJT that did "succeed in providing a satisfactory discrimination between artists and controls" (Eysenck & Castle, 1971, p. 392). In this study, we administered a 22-item subset, an extension of the 13-item subset (Eysenck, 1972b).

While Graves (1948) based his construction of the items on "principles of aesthetic order--unity, dominance, variety, balance, continuity, symmetry, proportion, and rhythm" (p. 2), an inspection of the 22-item subset indicates they simply contrast symmetric and asymmetric designs, with the keyed choice being the asymmetric design. Figure 1 presents an item similar to those in the 22-item subset. As can be seen on the left, a vertical line through the horizontal lines results in a design that is slightly off center unlike the design on the right.

Visual Designs Test. The conceptual bases for the VDT are principles from visual perception and Gestalt psychology. In studies of perception, researchers have found that viewers organize visual stimuli into sub-units and over repeated exposures build up a percept (Haber & Hershenson, 1965) that corresponds to a verbal concept. This extraction of information from a visual image occurs instantaneously (Brighouse, 1939a; Kellet, 1939; Külpe, 1903); Haber and Hershenson (1965) found that the process occurs with exposures as brief as five milliseconds, although the time necessary for forming a percept depends on the content of the image. Gestalt psychologists have argued against an elemental approach such as this to perception, although they too suggest that a viewer operates on an image by

Figure 1

Sample Item from Design Judgment Test

Examinees select the design they prefer.
The keyed choice is the design on the left.

extracting its simplest structure (Koffka, 1935; Köhler, 1920). This principle of information extraction guided the development of the VDT, in which a visual image is conceptualized as consisting of a cell structure in which groups of cells function together to form patterns and designs. As a consequence, any manipulation of the cells that underlie an image changes the overall pattern, and possibly the preference for it. When one systematically constructs designs that differ in characteristics known to distinguish between artists and nonartists, these designs, when used as test items, should be effective in identifying persons with preferences similar to artists. On the basis of considerable previous research, complexity and order were the characteristics chosen for manipulation in the VDT.

This theoretical perspective was put into operation by means of a systematic design-construction program consisting of rules for the assignment of visual elements to randomly chosen locations within 8" x 11" visual fields (Technical Report 1988-1). Although programmed artwork (using in part explicitly random factors) has been developed by others (Attneave, 1956, 1957, 1959; Dorfman, 1965; Noll, 1966, 1972), researchers have not simultaneously manipulated complexity and order nor used designs of this type in individual-difference measures.

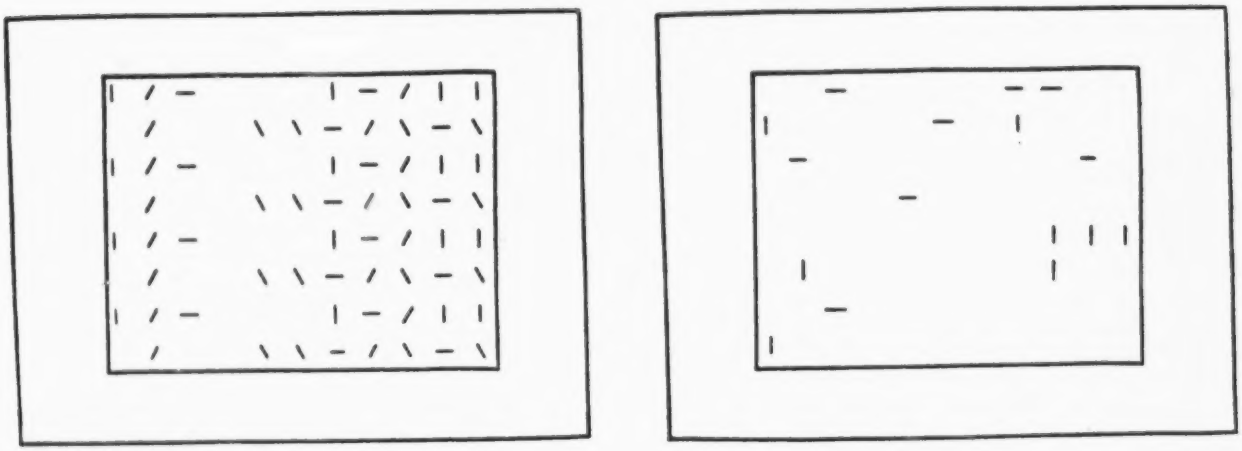
For the VDT designs, the total number of elements assigned to a visual field (similar to Birkhoff's operational definition of complexity, i.e, the number of sides of a polygon; see Birkhoff, 1956, for a discussion) determined the approximate complexity of a design, while manipulation of the repetition of a pattern in a design by 50 or 100% controlled the order. Designs with a great deal of repetition were considered to possess a high degree of order, while the absence of repetition resulted in totally random patterns. Our manipulation of order differs from the operational definition of other researchers (Garner, 1970) because our method does not result in a symmetrical pattern or mirror image for

patterns with a high degree of order. Instead we follow a model presented by Attneave (1959) in which order is operationally defined by a pattern of repetition across a design. According to Attneave, this method should produce differences in persons' liking for a pattern, and thus we expect it to be useful in distinguishing between artists and nonartists. The complete item-construction specifications are given in Technical Report 1988-1 (pp. 36-37). Figure 2 presents two visual designs used in a VDT test item.

In 1987, the Atlanta and Dallas testing centers presented two sets of 45 VDT designs each to examinees. The examinees rated the attractiveness of each design on a scale of 1 to 5. The results indicated that greater complexity was associated with

Figure 2

Sample Item from Visual Designs Test



Examinees select the design they prefer. The design on the left is constructed to be more-complex.

higher attractiveness ratings by laypersons for the designs, while order had a systematic effect only in its interaction with complexity.

Following on the rating-scale results, Bezruczko formed a forced-choice test presenting 84 pairs of designs (see Research Memorandum 1989-3) that contrast higher levels of complexity with lower levels of complexity. The respective levels of order were not controlled and thus vary from item to item. Examinees are instructed to select the design from each pair that they prefer. This version of the test, henceforth referred to as the VDT, was administered to the examinees in this study.

Proportion Appraisal (Worksample 235 C). In this study we used the most recent version of PA, Worksample 235 C (Technical Report 44). This test consists of 50 items, each containing three simple geometric figures differing only in their proportions. For one figure in each item, the ratio of the width to the length is .50 (1:2); for another, .67 (2:3); and for the third, .75 (3:4). On one item (item 50), we were unable to precisely identify the proportions. The ratio of .67 corresponds the closest of the three to the ratio of the golden section, a favored art concept, and tends to characterize the most-preferred figures on the test (Technical Report 99).

PA test directions instruct the examinee to select two figures from the three presented for each item, namely, the examinee's choices for best figure and worst figure. In the past, scoring of item responses was based on the ". . . consensus of opinion of the general population. . ." (O'Connor, 1940, p. 113), or the correspondence between an examinee's responses and the most-common responses to items. According to O'Connor (1940), "The test selects men and women whose judgment of proportions conforms with that of humanity" (O'Connor, 1940, p. 111). This method resulted in some items with ratios of .67 as the keyed figure, while for other items .50 or .75 was the keyed choice and for other items the scoring rationale was unclear.

During its use in the Foundation, Proportion Appraisal has been found to be moderately reliable (.70) and correlated -.43 with the Design Judgment Test (Technical Report 792). Earlier Foundation research indicated that artists score low on the test (Technical Report 792), and Foundation writers have speculated that artists tend to prefer the elongated forms (ratio of .50) (Trembly, 1974). This would suggest that persons who prefer more-elongated figures on PA will tend to prefer the asymmetric designs on the DJT.

In order to clarify the influence of proportions on preference, we created separate keys for .67 and consensus scoring. In the first of these scoring methods, the figure with the .67 proportions was keyed as having the best shape. Initially, examinees were given one point for each item on which they chose the .67 figure for the best. Then we gave examinees a point each time they chose the .50 figure as the worst shape (the .50 figures tended to be the most-common choices for worst shape). We refer to this method of scoring, with .67 for best shape and .50 for worst shape, as ".67" scoring.

For the second scoring method, we relied on the obtained preference responses for each item. Rather than keying the .67 figure as the best shape and .50 as the worst shape on all items, we keyed whichever figures were most often chosen on each item as the best shape and the worst shape. Thus, examinees received one point for every choice for best that matched the group choice and one point for each choice for worst shape that matched the group

choice. For this scoring method, several items were not scored for best or worst because there was no clear choice (i.e., no response given by more than 40% of examinees). We refer to this scoring method as consensus of opinion scoring (or simply consensus scoring). Figure 3 presents a sample PA item.

Figure 3

Sample Item from Proportion Appraisal

.50 .75 .67

Which figure has the BEST shape?
Which figure has the WORST shape?

Visual Aesthetic Sensitivity Test (Götz, 1981). This test presents examinees with 50 pairs of nonrepresentational designs artistically rendered by an internationally renowned painter. The test developers produced the designs in each pair in such a manner that one of them is considered to have better harmony and balance, and the examinee is instructed to identify that design in each pair, as specified in the test instructions:

> Each pair consists of two quite similar pictures, one of which is superior from the point of view of design; it is more harmonious, better balanced, and better adapted in the way the elements are ordered and in the way the lines are drawn. . . . Your task will be to discover . . . which in each pair is the better design. (Götz, 1981, instruction sheet)

The test developer, K. O. Götz, argues that the keyed designs on the VAST are objectively "better" than their nonkeyed counterparts in terms of balance and harmony and that the test's validity is therefore independent of the styles and fashions of any particular point in time. In support of this position, he reports a review of the designs by eight professional artists (Götz et al., 1979). For all 50 items in the published version

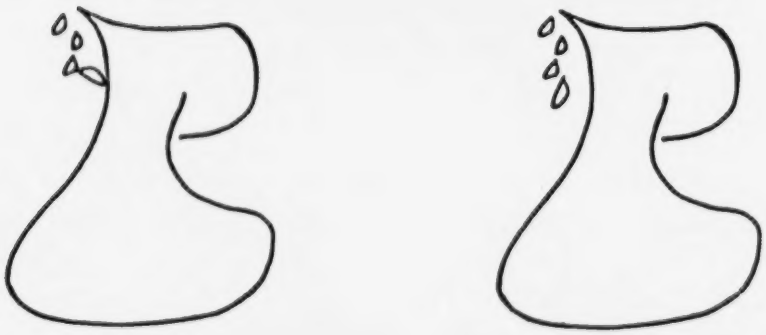
of the test, the eight artists agreed that the keyed design was better balanced and more harmonious than the nonkeyed design.

Examinees receive one point for each item on which their choice of the better design (in terms of balance and harmony) agrees with the key. Thus, examinees' abilities to identify balance and harmony are evaluated against the standards set by the test developer. It should be noted that the other artistic judgment tests studied here, in contrast, rely on personal reports of preference and do not ask examinees to make objective judgments of value.

Visually, the VAST comprises two types of items. The first type of items contains simple concrete designs that generally appear as full shapes, oftentimes with rounded contours, although a few have sharp points. The designs in these items generally have little detail. The second group contains designs with broad sweeping brush strokes, sometimes producing an abstract circular motion. These designs are generally very complex, with an extensive amount of fine detail in the brush work. Figure 4 presents an item similar to the simple concrete type from the VAST.

Figure 4

Item Similar to the Items on the Visual Aesthetic Sensitivity Test



The design on the right is the better of the two designs in terms of balance and harmony.

According to Götz et al. (1979), photographic transparent slides of 42 of the 50 test items have been administered to groups of nonartist adults and found to have a reliability of .84. The present study used a printed version of the VAST, in which the designs are reproduced as duotone (black and white)

prints. Visual differences between projected slide images and two-dimensional print copy may influence the test's psychometric properties, although to what extent has not yet been explored. In particular, the highly detailed features of the brush-stroked items may not be as apparent in the printed version as in the slide version, thus influencing preferences. Given the convenience of administering printed items, the analysis of the reliability of the printed form of the VAST is an important aspect of this study for the Foundation.

Procedures

The artistic judgment tests were administered by means of written instructions at the end of testing sessions with the standard Foundation battery. The tests were administered in this order: VDT, DJT, PA, VAST. In all cases, a test that was started in one session was completed in that session. None of the tests were timed, and the standard directions for administration were used. The items were presented in booklets, and examinees marked their responses on separate answer sheets.

Analyses

The purpose of the analyses in this study was to examine the internal structure of the four artistic judgment tests. For each test, the analyses consisted of several stages. First, we computed item statistics along with the overall alpha reliability for the test. In this same stage, we also performed a factor analysis of the test items. In a second stage, we calculated sex and age effects. In the final stage, we examined the relationships among the tests. We describe these analyses in further detail in the remainder of this section.

Unless otherwise stated, the SPSS/PC (Version 2.0; Norusis, 1988a), SPSS/PC+ Advanced Statistics (Version 2.0; Norusis, 1988b), MSCALE (Version 2.0; Wright, Congdon, & Rossner, 1987), and BANK (Wright & Schulz, 1988) computer programs were used for these analyses.

Item Analyses

We computed values for the following statistics for the items in each of the tests: (a) p values (proportion of examinees selecting the keyed response) and Rasch (one-parameter logistic) difficulty estimates, (b) item-total correlations, and (c) Rasch item fit statistics. We also calculated the overall test reliabilities and performed item factor analyses for each test.

Item difficulty. The simplest indication of item difficulty is the p value, which is the proportion of persons who pass an item (i.e., their answer matches the keyed response). A p value of .90, for example, indicates that 90% of the examinees passed the item. P values provide a quick means of establishing the

distribution of abilities covered by a set of test items. Many items at the same p value indicate redundancy in the test design and may result in substandard test reliability. A related problem is a test with many items of extreme difficulty, either high or low.

Although item p values are useful for ranking the difficulty of test items, they have several limitations. One is that their characterization of difficulty is specific to the sample tested, and another is the unequal distances between units on the percentage scale (e.g., the distance between 50 and 60% correct is different from the distance between 85 and 95% correct). These problems are generally remedied by transforming the p value to some standardized deviation score. In this study, we transformed all p values to logits, as estimated by the Rasch measurement model. This operation expresses the item difficulties in equal-interval scale units and also addresses the problem of sample dependency. Since a Rasch difficulty estimate is an inferential statistic, item difficulty can be interpreted independently of the group tested, thus permitting valid generalizations beyond the immediate sample tested (Wright & Stone, 1979). Transformed logit difficulty values for the items in the four artistic judgment tests appear in Research Memorandum 1989-2.

Item-total correlation. The item-total correlation indicates the consistency between examinees' responses to an item and their total test scores. In general, the examinees who answer a given item correctly can be expected to have higher total test scores than examinees who answer the item incorrectly, so that effective test items have positive item-total correlations. For each item, then, we calculated the correlation between the item and the total test score minus that item. (This value is sometimes referred to as the corrected item-total correlation.) We considered item-total correlations of at least .30 to indicate adequate item functioning.

Item fit. A fundamental operation in test development is establishing whether test items work together in operationally defining a construct. Specifically, do the persons who passed a given item correspond to the persons with higher ability estimates on the overall scale? To answer this question, we examined the standardized weighted infit values of the Rasch fit t statistic (for a discussion, see Wright & Masters, 1982) for the items on the four tests. These values are based on the difference between the expected response of an examinee for an item as estimated from his total test score and the observed response that the examinee marks on the answer sheet. When the response for a person to an item departs significantly from statistical expectation, the fit t statistic increases in magnitude.

For these analyses, we used a fit criterion of 3.00. Values of the fit t statistic that exceed this criterion may indicate responses to items that do not meet fundamental requirements for linear measurement. A limitation of this procedure, however, is the influence of our large sample size on the value of the fit statistic. In view of this, we were cautious in our interpretation of fit t values, using them only as evidence that specific items may not be consistent with the rest of the items in a test.

Reliability. Since none of the tests were timed, the reliability of item responses was estimated using coefficient alpha as shown below:

$$r_{tt} = \left(\frac{k}{k-1}\right)\left(\frac{s_y^2 - \Sigma s_i^2}{s_y^2}\right)$$

where r_{tt} is coefficient alpha, k is the number of items in the worksample, s_y^2 is the variance of the total score, and s_i^2 is the variance of an item score (Nunnally, 1967). The Foundation has established a standard of .80 for acceptability in terms of reliability in order for a test to be included in the standard battery of aptitude tests (Test Information Bulletin 1980-5).

Item factor analysis. We used the principal components method of factor extraction on each test to determine if a single factor adequately accounted for the variation in item scores. When it was necessary to extract multiple factors for a test, the varimax rotation of factors was used. The decision regarding the number of factors needed to represent item responses was based on the eigenvalues obtained in the factor analysis and the related percentage of total variance accounted for by each factor. Eigenvalues greater than 1.00 indicated factors of interest. In making these determinations, we also examined scree plots of the eigenvalues.

Analysis of Sex and Age Effects

Sex and age effects for each test were examined with an analysis of variance (ANOVA) of total test scores. The ANOVA was designed to answer three questions:

1. What is the main effect of sex on artistic judgment test scores when differences in age are controlled?
2. What is the main effect of age on artistic judgment test scores when scores are adjusted for differences in sex? and
3. Do sex and age have an interactive effect on scores?

For these analyses, we partitioned age into the following eleven groups: (a) 14-15 (b) 16-17, (c) 18-19, (d) 20-23, (e)

24-27, (f) 28-31, (g) 32-35, (h) 36-39, (i) 40-43, (j) 44-47, and (k) 48 and older. Sex was a second independent variable, and the resulting 2 x 11 design contained unequal sample sizes in the cells.

For these analyses, we used the regression approach to ANOVA. We assessed the size of each statistically significant main effect by calculating the value of the omega-squared (ω^2) statistic (Cohen & Cohen, 1975). Omega-squared is the proportion of variance in the score that is attributable to the given effect. Note that the square root of the omega-squared value indicates the size of a correlation with the same effect size. For example, an omega-squared of .16, omega = .4, indicates an effect size equivalent to the effect of a correlation of .4.

Analysis of the Artistic Judgment Tests

After examining the internal psychometric characteristics of the four tests, we examined the relationships between the total scores for the tests. First, we used Pearson product-moment correlations to measure the magnitude of association between the total scores for the four tests. Next, we performed an exploratory principal-axes factor analysis on the correlation matrix for the four tests. The factor analysis determined whether the relationships among the tests were sufficiently strong to indicate a general aptitude factor. In general, we considered factors with eigenvalues greater than 1.00 to indicate aptitude factors.

The results of these analyses are reported in the following section.

RESULTS

The Artistic Judgment Tests Considered Individually

Design Judgment Test

Descriptive statistics. Total test scores were unimodally distributed around the center of the scale, ranging from the test floor of 0 to the test ceiling of 22 (M = 10.8, SD = 5.9, Mdn. = 11.0). There was moderate skewing toward the high end of the scale (preference for asymmetry), and a floor effect occurred for 36 (2.18%) examinees and a ceiling effect for 41 (2.48%) examinees.

Internal structure. Values of the item statistics for the DJT appear in Table 1. The p values indicate that six of the items are moderately difficult, with more than 60% of the examinees choosing the non-keyed design. Two of the items were moderately easy, with over 70% of the examinees agreeing with the

key. Seven of the items were at or very close to the 50% value, which is the p value that corresponds to random guessing on these items.

Table 1

Item Statistics for the Design Judgment Test

Item[a] no.	Difficulty (p value)	Item-total corr.	Rasch infit value
4	.49	.16	16.71
6	.34	.49	-.35
8	.71	.52	-2.90
10	.53	.49	1.12
16	.51	.42	4.38
21	.60	.50	.22
25	.69	.46	.72
28	.55	.37	7.30
29	.58	.56	-2.91
42	.70	.37	4.14
44	.34	.44	2.69
47	.66	.54	-2.86
48	.57	.59	-5.33
50	.27	.53	-3.61
51	.34	.49	-.55
53	.37	.56	-4.44
57	.47	.55	-2.65
59	.40	.61	-6.37
61	.40	.62	-7.51
66	.46	.56	-2.54
69	.50	.53	-1.21
72	.37	.52	-1.82

Note. N = 1,650. Items are scored in the direction of preference for asymmetry.

[a]The item numbers correspond to the position of these items in the original DJT test of 90 items.

The alpha reliability for the 22-item version of the DJT was relatively high, .89. Item-total correlations ranged from .16 to .62 with a mean of .49. The highest item-total correlations were for Items 29, 48, 53, 59, 61, and 66. Visual inspection indicates that among these items, Items 61 and 66 involve choices between pairs of straight lines. The lowest item-total correlation was for Item 4.

Rasch infit values for four of the items (4, 16, 28, and 42) were above the criterion of 3.00, indicating poor fit, but with a sample size of 1,650 their misfit may not be substantial. The large infit value for Item 4, however, does suggest that responses for that item differ from the expectations of the measurement model. An inspection of the test indicates that Item 4's position as the first item in the test may be a source of its problem. Consideration should be given to either dropping the item from the subset or changing it to an unscored sample item. Dropping or not scoring the item will raise the alpha reliability to .90.

A principal components factor analysis yielded five factors with eigenvalues greater than or equal to 1.00. The values are 6.97, 1.43, 1.16, 1.05, and 1.03, respectively. The first factor accounted for 31.7% of the variance with the remaining factors accounting for a combined total of 21.3%. An examination of the scree plot indicated that a one-factor model is adequate for this set of items.

Sex and age effects. The analysis of variance of test scores by sex and age indicated significant main effects for sex and age for the DJT, F (1;1,439) = 70.97, $p < .001$ and F (6;1,439) = 3.34, $p < .05$, respectively. The interaction between sex and age was not significant.

Average scores for females and males were 12.15 and 9.58, indicating a higher preference by females for asymmetrical designs. Adjusted for age, the difference between males and females was .43 of a standard deviation. The proportion of variance accounted for by sex was not large ($\omega^2 = .05$), comparable to the effect of a correlation of .22. The largest difference between age group means was .36 standard deviation units. After controlling for sex, the effect of age accounted for 1% of the variance in test scores ($\omega^2 = .01$), comparable to a correlation of .11.

The age curve for the DJT indicates a modest but continuous increase in test scores through the twenties, then a plateau in the thirties, with the hint of a decline beginning at age 38.

Summary. The main findings with regard to the DJT are:

1. The reliability of the 22-item version is high, .89. The test is easy to administer, and the time necessary for completion is brief.

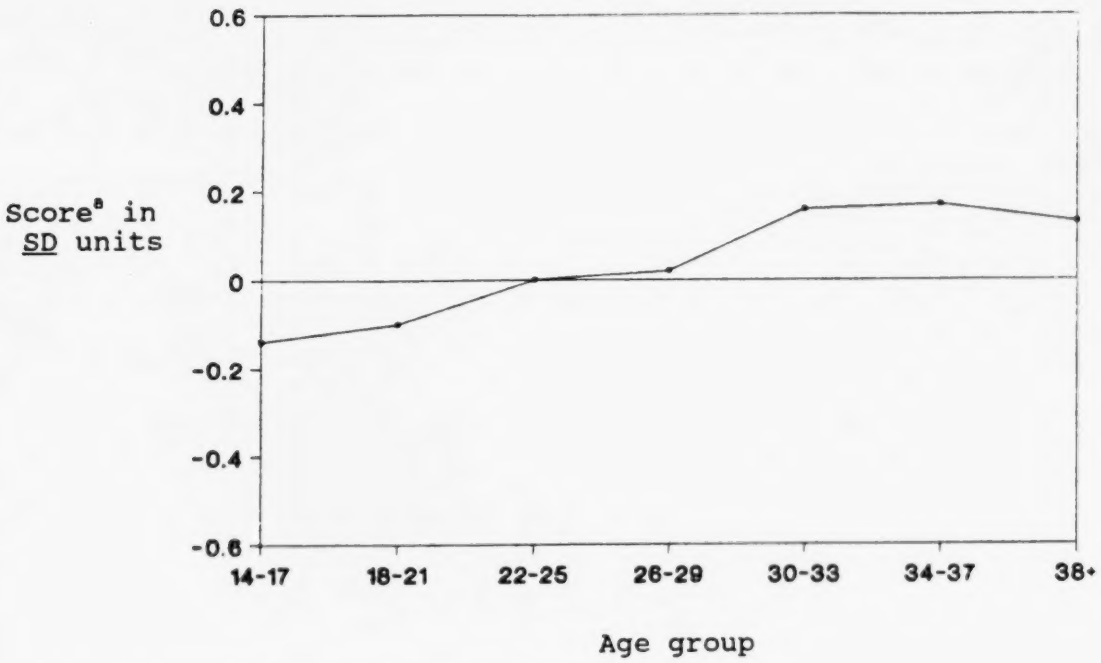
2. The DJT has mild floor and ceiling effects that appear to have little effect on measurement precision for the Foundation population.

3. Item 4 should be reviewed. Its item-total correlation is low, and its Rasch infit value is large.

4. The sex effect was moderate, with a stronger preference by females for asymmetrical designs. The age effect was mild, showing increasing scores from the mid-teens through the twenties and then a plateau in the thirties.

Figure 5

Age Curve for Design Judgment Test



Note. Plotted values are not smoothed.

[a]Adjusted for sex.

Visual Designs Test

Factor analysis. Because the VDT is a new test, we performed an exploratory principal components factor analysis to determine the number of dimensions it measures. This type of analysis is also useful for exploring content interpretations for a new test. In the principal components analysis, the eigenvalues for the first two factors were 18.16 and 10.25, accounting for 21.6% and 12.2% of the variance, respectively. The third factor was substantially smaller (eigenvalue = 3.02). When the three-factor solution was rotated to the varimax criterion, only two factors were clearly interpretable, and so the two-factor solution was retained.

As described in the Method section, the VDT consists of pairs of designs that contrast lower levels of complexity with higher levels. Visual inspection of the items loading on the first factor indicated that they involve, as might be expected, straightforward comparisons between more- and less-complex designs. We named this factor Simplicity.

Although the second factor also contains items with low- and high-complexity designs, the nature of the contrast differs from that found on the Simplicity factor. In a typical item for the second factor, the more-complex design tends to have elements grouped together and pointing in different directions, the result of the combination of complexity and redundancy specifically used to construct these items. This pattern in the more-complex design differs from the less-complex design in which all the elements are pointed in the same direction, oftentimes ordered in rows. The pattern in the more-complex design produces a somewhat disordered and agitated image that may be unsettling or disturbing to some examinees, in contrast to the less-complex design, which appears less disorganized and more uniform. We speculate that some examinees choose the lower-complexity designs as a function of avoidance of the more-aversive higher-complexity design rather than as attraction to simplicity.

Our examination of the item-construction specifications for the lower-complexity designs on the second factor indicates that in all cases they share the characteristics of only one type of element in a design, the highest level of redundancy, and the maximum number of panels possible per design. As a result, although the underlying generating process for the less-complex design is identical to that for the more-complex items, this particular set of item-construction specifications forces the assignment of elements to a design into a spatially dispersed pattern with visual characteristics of order and uniformity, and thus we name this factor Uniformity.

Descriptive statistics. To reiterate, the factor analysis indicated that the VDT measures primarily two variables. We formed scales for each of these variables consisting of the items with loadings greater than .50 on the corresponding factors in the rotated factor solution. Descriptive statistics for these scales appear in Table 2. The Simplicity scale contains 34 items, and the Uniformity scale contains 13 items.

The scores for Simplicity are unimodally distributed and moderately skewed toward higher complexity. Uniformity is also unimodally distributed but shows little skewing. Because Simplicity contains more items than Uniformity, it shows greater dispersion in its scores, with a standard deviation of almost twice the standard deviation of Uniformity. Floor effects were noted for 13 (<1%), and ceiling effects for 12 (<1%) of the examinees on the Simplicity factor. On Uniformity, floor effects were found for 78 (4.7%) and ceiling effects for 171 (10.3%) of the examinees.

Table 2

Descriptive Statistics for Visual Designs Test

Scale	No. items	M	SD	Mdn.	Range	Alpha
Simplicity	34	11.98	9.61	10.0	0-34	.95
Uniformity	13	7.46	4.07	8.0	0-13	.88

Note. The sample sizes for Simplicity and Uniformity were 1,654 and 1,658, respectively.

Internal structure. The item-total correlations for Simplicity ranged from .41 to .65 and for Uniformity from .41 to .71 with averages of .34 and .56, respectively. Reliabilities for both factors are high, .95 and .88, respectively. These values indicate the factors are effective in discriminating between persons with different preferences for designs. Item statistics for the Simplicity items appear in Table 3 and for Uniformity in Table 4.

On the Simplicity factor, the distribution of p values indicate that when presented with a choice between lesser or higher complexity, the simpler choice is less-often preferred. Only three of the items were answered in the keyed direction (Simplicity) by more than 50% of the sample. While the Uniformity factor is also scored in the direction of simpler designs, its distribution of p values differs noticeably from Simplicity. On Uniformity, over 50% of the examinees chose the keyed (more-uniform) design on 11 (84%) of the 13 items.

The Rasch infit values indicate that the items for both factors fit the modeled expectations for linear measurement. Only Items 3, 29, and 70 exceeded the infit criterion of 3.00 on Simplicity, although the sample size of 1,654 encourages caution in concluding that they misfit the model. The Rasch infits of Uniformity similarly indicates that items generally fit the measurement model, with the exception of Items 5, 37, and 79.

The value for Item 5 appears large relative to this set of items, and it should be investigated further. These results indicate that the items for these scales generally meet the requirements of a measurement model, and thus their transformation into linear measures is quantitatively valid.

As described in the Method section, the VDT is based on the manipulation of two factors, complexity and order. The factor

Table 3

Item Statistics for the Simplicity Factor on the Visual Designs Test

Item no.	Difficulty (p value)	Item-total corr.	Factor loading	Rasch infit value
1	.25	.48	.46	2.07
2	.18	.49	.51	-.15
3	.41	.48	.47	5.46
4	.36	.59	.54	-.88
6	.35	.55	.50	1.71
7	.25	.58	.61	-2.26
8	.28	.64	.59	-3.06
9	.41	.58	.54	.47
10	.40	.57	.57	1.00
11	.30	.53	.51	1.27
12	.28	.63	.49	-2.87
24	.19	.41	.50	2.53
25	.38	.65	.67	-2.74
29	.39	.51	.46	4.19
30	.37	.63	.64	-3.31
38	.32	.65	.70	-4.60
42	.44	.65	.66	-2.54
43	.28	.64	.65	-4.58
44	.31	.50	.55	2.34
45	.21	.44	.46	2.32
46	.29	.63	.67	-4.64
48	.40	.65	.67	-4.06
49	.38	.61	.63	-2.55
51	.34	.56	.55	.88
52	.50	.60	.59	-1.74
53	.32	.55	.57	.94
66	.23	.47	.55	1.70
67	.60	.52	.53	2.05
70	.54	.48	.52	4.44
71	.44	.55	.50	1.68
72	.44	.61	.59	-1.52
80	.35	.61	.66	-1.53
83	.39	.56	.61	.52
84	.30	.57	.57	-.70

Note. N = 1,654. Items are scored in the direction of preference for simplicity.

Table 4

Item Statistics for the Uniformity Factor on the Visual Designs Test

Item no.	Difficulty (p value)	Item-total corr.	Factor loading	Rasch infit value
5	.47	.41	.45	9.0
15	.55	.58	.64	-.5
16	.42	.52	.59	1.9
37	.65	.50	.52	3.2
56	.58	.57	.60	-1.1
57	.73	.61	.62	-3.5
58	.53	.61	.62	-1.8
59	.53	.71	.74	-9.5
61	.57	.58	.63	-.7
64	.63	.65	.66	-5.7
65	.58	.64	.66	-4.5
78	.50	.54	.58	2.0
79	.71	.43	.44	6.2

Note. N = 1,658. Items scored in the direction of preference for uniformity.

analysis, as we expected, indicated a powerful influence of complexity on Factor 1 (Simplicity) items, and a somewhat different effect on Factor 2 (Uniformity). Since the influence of order (manipulated in the form of redundancy) on preference was not apparent from these analyses, we formed a third scale consisting of items that differ in redundancy with complexity held constant. The more-redundant design was the keyed choice, so that the total scores on this scale indicate the preference for higher order. According to these criteria, we formed a 13-item scale on which total scores ranged from 1 to 11 with a mean of 6.5 (SD = 1.5, Mdn = 6). When we computed an alpha coefficient for the items, however, the scores were not reliable, and we did not perform any further analyses.

Sex and age effects. The analysis of variance indicated a significant main effect of age on Simplicity: $F(6;1,439) = 11.07$, $p < .001$. The largest difference between age groups was .65 standard deviation unit, and age effects accounted for a modest proportion of the variance ($w^2 = .04$), equivalent to the effect of a correlation of .21. The effects for sex and the interaction between sex and age were not significant.

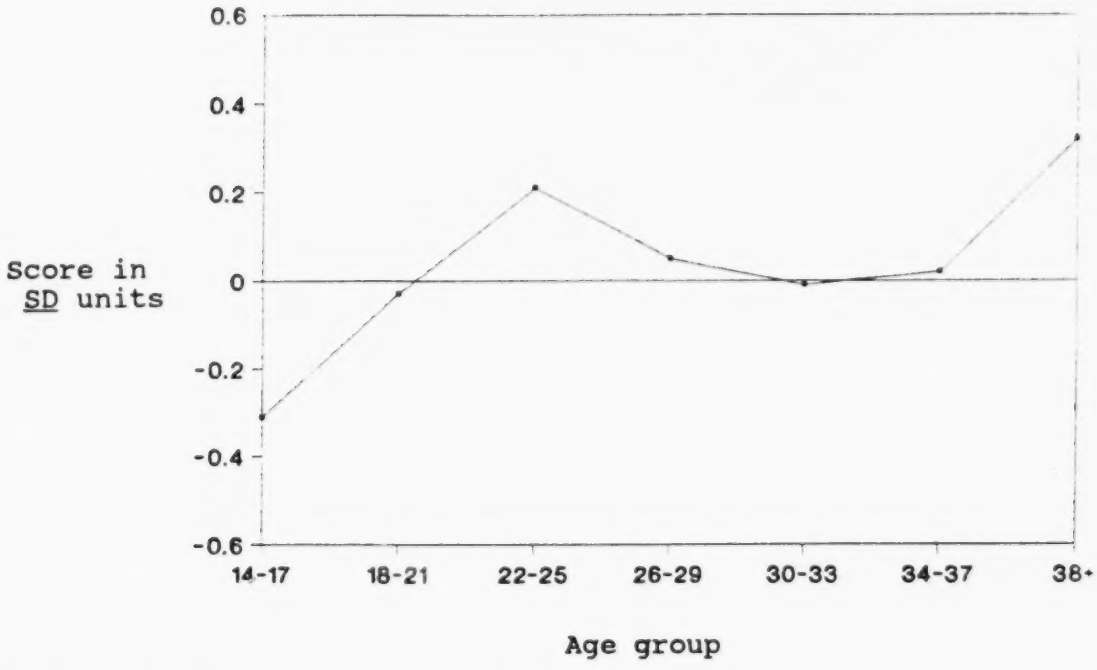
On Uniformity, the analysis of variance yielded significant main effects for both sex and age, $F(1;1,439) = 4.47$, $p < .05$; $F(6;1,439) = 2.74$, $p < .05$; respectively. The interaction between sex and age was not significant.

Males scored significantly higher on Uniformity ($M = 7.71$) than females ($M = 7.26$), although when controlled for age, the effect was small, .11 standard deviation units, accounting for a negligible proportion of the variance ($w^2 = .004$), the equivalent of a correlation of .06. The largest difference between age groups on Uniformity was .29 standard deviation units, and age accounted for only a small proportion of the variance ($w^2 = .012$), equivalent to a correlation of .11.

The age curves for Simplicity and Uniformity are presented in Figure 6 and 7, respectively. They indicate similar patterns of steadily increasing scores until the mid-twenties, a modest decline into the thirties, and then an increase for the 38+ group. The effects from age 22 on are small and may represent chance deviations around an overall plateau.

Figure 6

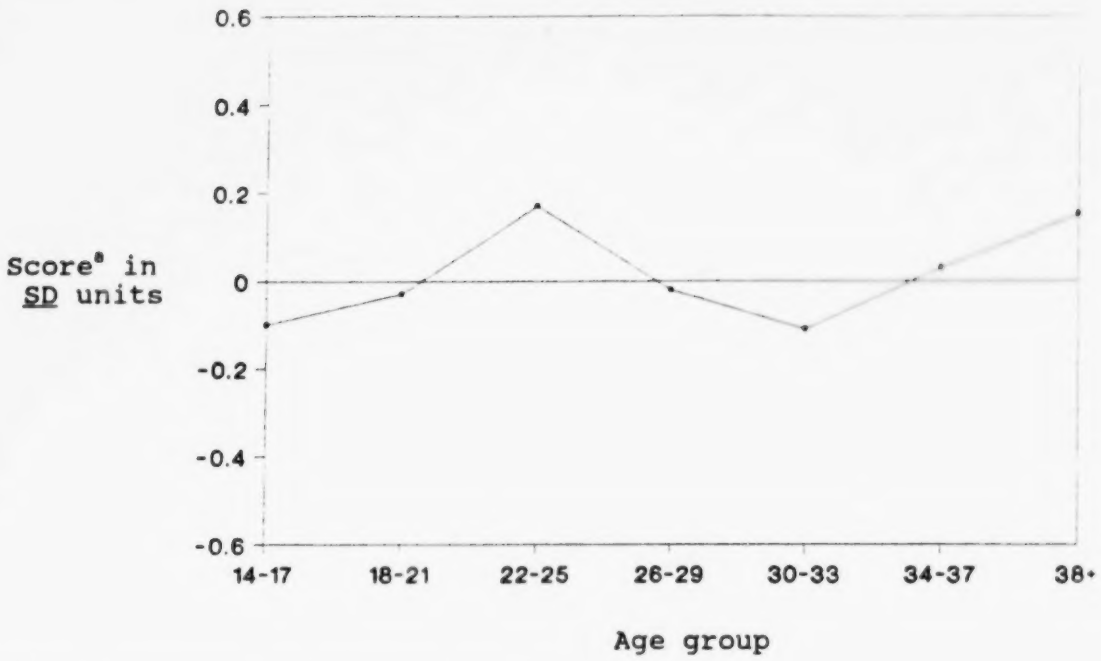
Age Curve for Simplicity



Note. Plotted values are not smoothed.

Figure 7

Age Curve for Uniformity



Note. Plotted values are not smoothed.

[a]Adjusted for sex.

Summary.

1. A principal components factor analysis indicated two factors for the VDT, with eigenvalues of 18.2 and 10.3, respectively. The two factors were called Simplicity and Uniformity, respectively.

2. Reliabilities for scales formed from the two factors were .95 and .88, respectively.

3. Ceiling effects and floor effects for two VDT factors did not appear to diminish the precision or stability of scores.

4. The p values for the VDT were noticeably lower for items on Simplicity than on Uniformity. Only three (9%) of the items on the factor called Simplicity were chosen by more than 50% of the examinees. On the factor called Uniformity, 11 items (85%) were chosen by more than 50% of the examinees.

5. Rasch infit statistics for Simplicity indicated three items (3, 29, and 70) with infit values of 5.46, 4.19, and 4.44,

respectively. For Uniformity, infit values for three items (5, 37, and 79) were 9.0, 3.2, and 6.2, respectively. In future administrations of these scales, their infit values should be examined to determine whether these items consistently exceed the fit criterion for the measurement model.

6. Items varying in order (redundancy) do not form a reliable scale.

7. A very small sex effect was found for Uniformity, indicating a slight preference of males for more-uniform patterns. Age effects for both Simplicity and Uniformity were moderate to small, accounting for a small proportion of the score variance. Both Simplicity and Uniformity scores were characterized by age curves that increase into the twenties, resemble a plateau somewhat through the thirties, and then increase slightly after the late thirties.

Proportion Appraisal

Descriptive statistics. Descriptive statistics for PA using the two scoring methods, consensus and .67, appear in Table 5. The distributions of total scores for consensus and .67 were similar. They were unimodally centered near the middle of the test with slightly positive skewness for consensus scoring. Neither of the scoring methods showed ceiling or floor effects.

Table 5

Descriptive Statistics for Proportion Appraisal

Key	No. items	*M*	*SD*	Mdn	Range	Alpha
Consensus	40	38.88	9.00	40	8-60	.76
.67	45	36.66	8.97	36	9-69	.78

Note. The *N*s for consensus and .67 scoring are 1,598 and 1,625, respectively.

The proportion of examinees choosing each shape for an item appears in Table 6. The average proportion of examinees that chose shapes with the ratio of .75, .67, and .50 was .44, .42, and .34, respectively. In terms of the choices for the worst shape, the average proportion of examinees was .40, .17, and .62, respectively. Thus the near-golden-section (.67) and .75 shapes

Table 6

Response Frequencies for Proportion Appraisal

Item no.	WORST .50	WORST .67	WORST .75	BEST .50	BEST .67	BEST .75
1	46	19	35	31	30	39
2	44	17	39	19	44	39
3	76	6	18	10	47	43
4	42	30	28	40	23	38
5	65	15	20	16	35	49
6	36	13	51	38	37	25
7	25	19	56	39	35	26
8	78	11	11	10	22	69
9	33	12	56	42	37	20
10	65	14	21	16	67	17
11	57	11	32	25	38	37
12	59	8	34	27	34	40
13	49	8	43	34	35	31
14	44	12	44	27	42	31
15	55	8	37	28	35	37
16	27	18	55	53	30	17
17	46	25	29	31	31	38
18	32	15	53	45	31	25
19	59	11	30	25	30	45
20	49	9	43	30	42	28
21	56	8	36	22	42	36
22	37	15	48	46	22	32
23	59	20	22	25	19	57
24	43	16	41	35	33	32
25	71	14	14	11	27	62
26	50	36	15	24	15	61
27	74	13	13	14	29	57
28	58	13	29	22	34	44
29	43	17	40	40	25	36
30	43	11	46	34	41	25
31	19	18	63	58	24	17
32	41	25	34	41	25	34
33	80	6	14	10	27	63
34	44	14	4	33	37	31
35	74	5	20	13	34	53
36	37	16	46	41	34	25
37	39	48	13	35	26	40
38	66	9	25	18	35	47
39	31	11	58	51	28	21
40	59	14	27	25	39	36
41	68	8	23	19	57	25
42	64	15	20	17	30	53
43	41	8	52	38	43	20
44	76	5	19	12	48	40
45	63	8	29	16	52	33
46	65	6	29	19	45	36
47	48	10	42	38	30	32
48	42	21	37	30	36	34
49	50	8	42	29	43	28
50	--	No Measure	--	--	No Measure	--

Note. Frequencies represented as proportions of examinees choosing each response. N = 1,660. Decimals omitted. Item 50 was deleted because its measured ratios were not consistent with the other items.

were very close as choices for the best shape, while .50 was clearly the shape that examinees found the worst.

Internal structure. The keyed responses, item-total correlations, and Rasch item fit values for consensus and .67 scoring are shown in Tables 7 and 8. For consensus scoring, the average item-total correlation is .24 with a range of .01 to .49. For .67 scoring, the item-total correlations range from .01 to .44 with an average of .24. The alpha reliabilities for consensus scoring and .67 scoring were .76 and .78, respectively.

The Rasch fit values tend to parallel the item-total correlations, with low item-total correlations corresponding to larger infit values. When Proportion Appraisal was scored on the basis of consensus of opinion, eight items exceeded the fit criterion (Items 2, 4, 7, 18, 30, 36, 37, and 49), while none of the items misfit the measurement model when the items were scored for the .67 proportion.

We performed a principal components factor analysis on the PA items, first using the consensus scoring key and then the .67 key. When we used the consensus key, the analyses yielded six factors with eigenvalues greater than 1.00, with the eigenvalues for the first four being 4.74, 1.78, 1.71, and 1.42. Thus, there is one primary factor for the test with some residual covariation. Similarly, when we used the .67 key, we found 13 factors greater than 1.00, with eigenvalues of 4.67, 2.10, 1.95, and 1.64 for the first four, respectively. We reached the same conclusion that the test has one primary factor along with some residual covariation.

Sex and age effects. For PA with consensus scoring, the analysis of variance indicated significant main effects for both sex and age: $F(1;1,439) = 4.51$, $p < .05$, and $F(6;1,439) = 4.00$, $p < .001$. The sex by age interaction was not significant.

Females ($M = 39.49$) scored slightly higher than males ($M = 38.45$) when PA was scored for consensus, indicating a tendency by females to prefer proportions more consistent with the general consensus than males. After controlling for age, however, the sex effect in standard deviation units was only .11, accounting for a negligible proportion of the variance ($w^2 = .004$), which corresponds to a correlation of .06.

The largest difference between age groups was .30 standard deviation units, and the age effect accounted for 1.7% of the variance, equivalent to a correlation of .13.

For .67 scoring, the analysis of variance showed significant main effects for both sex and age, $F(1;1,439) = 4.81$, $p < .05$, and $F(6;1,439) = 5.97$, $p < .001$. The sex by age interaction was not significant.

Table 7

Item Statistics for Proportion Appraisal, Consensus Scoring

Item no.		Keyed value[a] Worst	Keyed value[a] Best	Item-total correlation	Rasch infit value
1		.50	--	.30	-1.98
2		.50	.67	.04	3.90
3		.50	.67	.23	1.24
4		.50	--	.01	3.87
5		.50	.75	.37	-3.03
6		.75	--	.17	-.02
7		.75	.50	.11	4.83
8		.50	.75	.49	-7.49
9		.75	.50	.19	2.68
10		.50	.67	.24	-1.36
11		.50	--	.43	-7.16
12		.50	.75	.35	-2.28
13	DEL'D	--	--	--	--
14	DEL'D	--	--	--	--
15		.50	--	.41	-5.68
16	DEL'D	--	--	--	--
17		.50	.75	.19	2.89
18		.75	.50	.10	3.72
19		.50	.75	.42	-4.54
20		--	.67	.10	1.55
21		.50	.67	.35	-1.23
22	DEL'D	--	--	--	--
23		.50	.75	.40	-5.30
24	DEL'D	--	--	--	--
25		.50	.75	.26	-.77
26		.50	.75	.21	1.25
27		.50	.75	.20	-.15
28	DEL'D	--	--	--	--
29		.50	--	.15	1.72
30		.75	.67	.10	3.37
31		.75	.50	.20	.60
32	DEL'D	--	--	--	--
33		.50	.75	.25	-2.31
34	DEL'D	--	--	--	--
35		.50	.75	.40	-4.48
36		.75	.50	.12	4.04
37		.67	.75	.09	3.80
38		.50	.75	.46	-5.40
39	DEL'D	--	--	--	--
40		.50	--	.05	1.62
41		.50	.67	.26	-.63
42		.50	.75	.47	-5.20
43		.75	.67	.18	2.31
44		.50	.67	.18	1.01
45		.50	.67	.30	-.89
46		.50	.67	.14	2.43
47	DEL'D	--	--	--	--
48		.50	--	.31	-1.75
49		.50	.67	.10	5.06
50[b]		--No measure--		.09	2.06

Note. N = 1,598.

[a]Deleted items indicate a consensus of opinion less than 40%.

[b]Although the measured ratios for Item 50 were not consistent with the other items, we were able to key it for preference and have included it in the analysis.

Table 4

Item Statistics for Proportion Appraisal, .67 Scoring

Item number		Item-total correlation	Rasch infit value
1		.26	-.86
2	DELETED	--	--
3		.27	-1.24
4		.13	-.42
5		.32	.33
6		.22	.04
7		.17	-.21
8		.23	-1.30
9		.17	1.25
10		.20	.24
11		.32	-2.64
12		.34	-2.99
13		.23	-.48
14	DELETED	--	--
15		.38	-1.14
16		.19	-1.05
17		.12	3.04
18		.19	-1.33
19		.41	-5.19
20		.24	-2.08
21		.36	-1.20
22		.29	-1.13
23		.32	-2.15
24		.13	1.39
25		.11	1.71
26		.13	1.77
27		.01	.42
28		.22	-.42
29		.26	-.86
30		.26	-.74
31		.08	.62
32		.30	-1.78
33		.20	.20
34		.21	.64
35		.35	-3.62
36		.24	1.50
37		.08	2.33
38		.44	-5.71
39		.22	-.04
40	DELETED	--	--
41	DELETED	--	--
42		.40	-4.47
43		.25	-.79
44		.14	2.37
45		.33	-2.58
46		.16	.41
47		.26	-2.92
48		.25	-.35
49		.25	2.78
50	DELETED[a]	--	No Measure --

Note. N = 1,625. Deleted items were removed from the scale because of low item-total correlations.

[a]Item 50 was deleted because its measured ratios were not consistent with the other items.

Females scored (M = 37.34) higher than males (M = 36.27) when PA was scored for the proportion, .67, indicating a slightly stronger preference by females for shapes that approximate the golden section. The sex effect, however, was only .11 standard deviations units and accounted for a negligible proportion of the variance (w^2 =.004).

The largest difference between age groups was .44 standard deviation units, and age accounted for 2.6% of the variance, equivalent to a correlation of .16. The age curve for consensus scoring in Figure 8 indicates a pattern of modestly increasing scores through the middle twenties and then a plateau through later adulthood. The age curve for .67 scoring in Figure 9 shows a fairly similar pattern, with a slight drop for the 30-33 age range.

Because neither scoring method showed a clear advantage in terms of internal psychometric qualities, a choice between them will have to be made on the basis of validation data (see later reports in this series). Using either scoring method, one would need to improve the test's reliability to some extent in order to reach the Foundation's standard of .80.

Summary.

1. Alpha reliabilities for consensus and .67 scoring were .76 and .78, respectively. Mean item-total correlations were .23 and .24, respectively.

2. Principal components factor analysis indicated that a one-factor solution is marginally adequate for both consensus and .67 scoring, with eigenvalues for the first factors of 3.11 and 4.67.

3. Overall, examinees tended to chose .75 as the best proportion and .50 as the worst proportion.

4. The near-golden-mean proportion, .67, was infrequently chosen as the best or the worst.

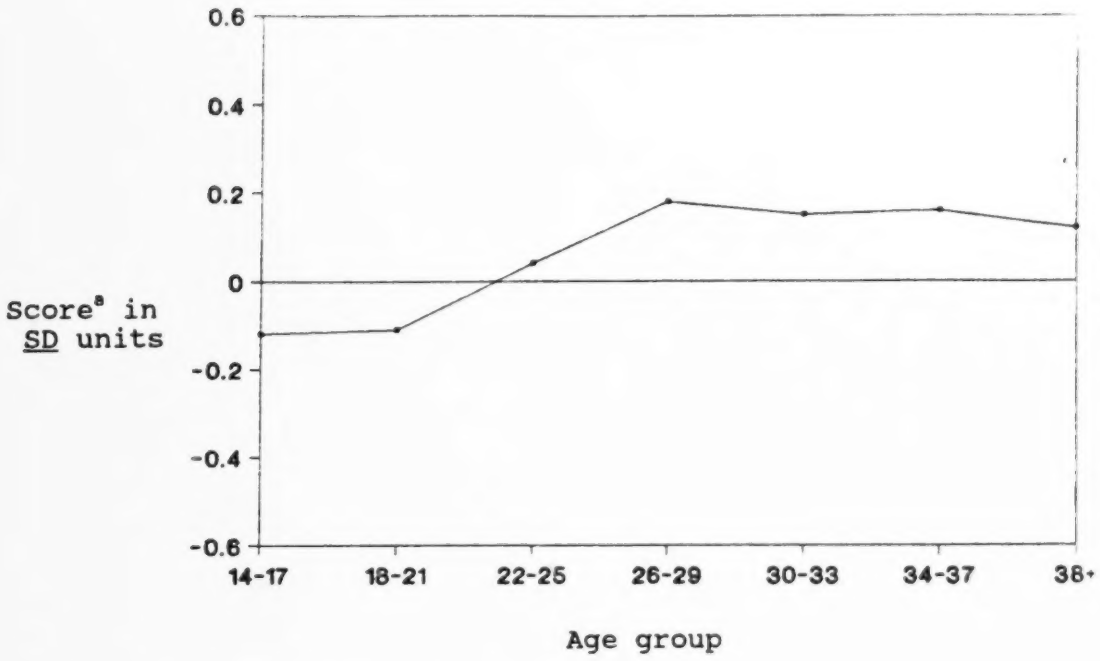
5. No ceiling or floor effects were noted for either scoring method.

6. Rasch infit statistics indicated that eight items exceeded the fit criterion of 3.00 when PA was scored for consensus. None of the items, when scored using .67, exceeded the infit criterion.

7. Sex and age effects for consensus scoring were modest: .11 and .30 standard deviation units, respectively. Sex and age effects for .67 scoring were .11 and .44 standard deviation units, respectively.

Figure 8

Age Curve for Proportion Appraisal, Consensus Scoring

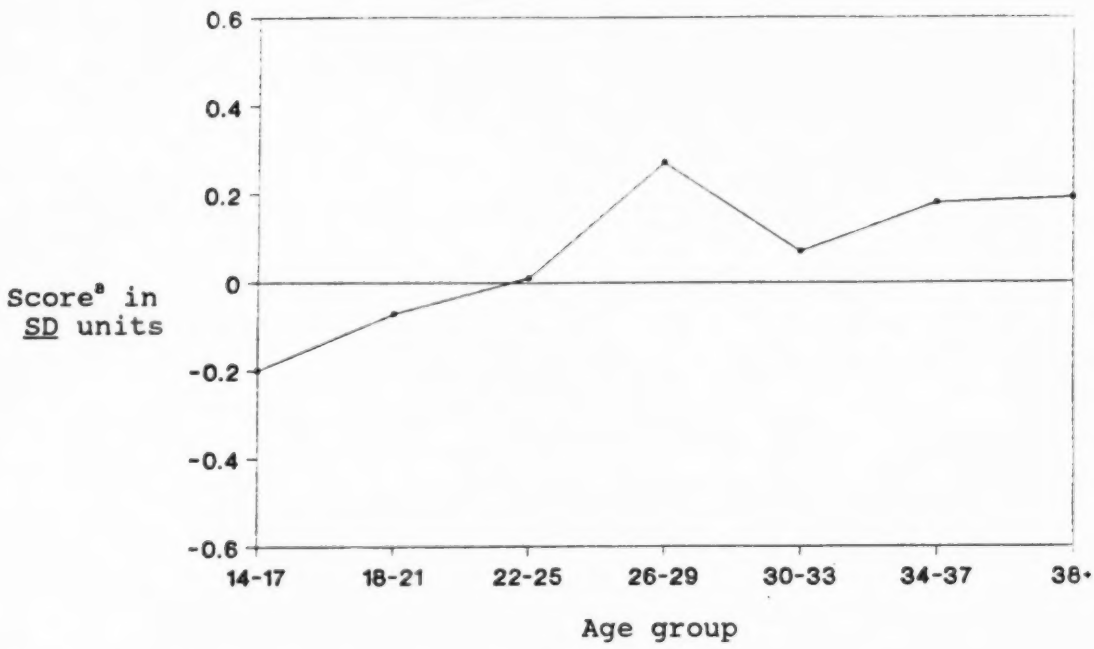


Note. Plotted values are not smoothed.

[a]Adjusted for sex.

Figure 9

Age Curve for Proportion Appraisal, .67 Scoring



Note. Plotted values are not smoothed.

[a]Adjusted for sex.

Visual Aesthetic Sensitivity Test

Descriptive statistics. Total test scores for the VAST ranged from 20 to 50 (M = 40.1, SD = 4.6, Mdn. = 41). Scores were symmetrically distributed around the center of 40, which represents 80% correct.

Internal structure. Item statistics for the VAST appear in Table 9. The p values for 20 of the items were equal to or greater than .85. Three of the items had p values equal to or greater than .95, and a ceiling effect on the test was observed for two persons. No floor effect was observed.

Two of the items (16 and 21) were passed by only 50 to 60% of the sample. Because 50% is the guessing level for the VAST items, consideration should be given to whether these items are performing effectively in the test.

Item-total correlations were generally low, ranging from .00 to .30 with a mean of .17. Only 13 of the items (3, 5, 7, 9, 12,

15, 24, 29, 38, 40, 43, 44, and 48) had item-total correlations over .20. These items tend to be the easier items (11 have p values > .86) and consist visually of simple, concrete designs characterized by round masses or solid angular pieces. Three of these items (5, 12, and 24), however, are from the brush-stroked group described in the Method section.

The alpha reliability for this version of the VAST was .66, which is low and indicates the test would need substantial improvement before it could be used for individual assessment at the Foundation.

To determine whether a subset of the VAST items may have better internal consistency than the total test, three groups of items were analyzed separately: (a) items with item-total correlations greater than .18, (b) items with item-total correlations greater than .20, and (c) another group for which the first author (Bezruczko) judged that the correct response is clearly consistent with the test directions regarding balance and harmony. The results of these analyses appear in Table 10. As the table indicates, although these subscales show improved average item quality compared to the overall scale (reflected in the higher item-total values), the reliabilities are less than for the full test. It appears that neither the full 50 VAST items nor any particular subset of items, in the form administered here (see later discussion), yields satisfactory measurement precision.

The item factor analysis for the VAST had to be handled in a special way. Because of computer limitations, the items were split into two subsets, odd-numbered and even-numbered, and a principal components factor analysis was run on each set of 25 items. For the odd-numbered items, factoring resulted in 10 factors with eigenvalues greater than 1.0, and the values for the first four were 2.33, 1.41, 1.29., and 1.23, respectively. The first factor accounted for 9.3% of the total variance, and the second and third factors accounted for 5.6% and 5.2%, respectively. None of the remaining factors accounted for more than 5% of the variance. For the even-numbered items, factor extraction likewise yielded 10 factors with eigenvalues greater than 1.0, with values of 2.13, 1.48, 1.37, and 1.24 for the first four, respectively. The total variance of the even-numbered items attributable to the first factor was 8.5%. The second and third factors accounted for 5.9% and 5.5%, respectively, of the total variance.

To ascertain if the first factor extracted for the odd-numbered items was equivalent to the first factor extracted for the even-numbered items, the two factor scores obtained for each examinee were correlated. The correlation between the odd and even factor scores was .53, which, given the alpha reliability of .66, indicates reasonable agreement between the two subsets.

Table 9

Item Statistics for Visual Aesthetic Sensitivity Test

Item no.	Difficulty (p value)	Item-total corr.	Rasch infit value
1	.85	.11	.51
2	.79	.15	.35
3	.89	.24	-.65
4	.63	.02	4.80
5	.82	.24	-.86
6	.75	.18	-.18
7	.78	.27	-1.35
8	.82	.17	-.02
9	.92	.29	-.83
10	.77	.15	.52
11	.89	.20	-.41
12	.74	.29	-2.08
13	.94	.14	-.08
14	.90	.14	-.09
15	.86	.27	-1.07
16	.55	.00	7.69
17	.62	.18	.38
18	.78	.16	.10
19	.64	.16	.66
20	.75	.18	-.16
21	.58	.08	3.74
22	.85	.13	.31
23	.80	.18	-.21
24	.70	.22	-1.01
25	.75	.08	1.71
26	.78	.09	1.26
27	.86	.13	.19
28	.88	.19	-.31
29	.86	.23	-.70
30	.73	.18	-.01
31	.81	.20	-.47
32	.84	.15	.07
33	.78	.12	.88
34	.84	.13	.27
35	.91	.20	-.38
36	.83	.15	.06
37	.88	.17	-.23
38	.87	.23	-.69
39	.87	.18	-.20
40	.89	.26	-.91
41	.75	.10	1.54
42	.70	.07	2.45
43	.95	.28	-.48
44	.97	.25	-.20
45	.92	.10	.17
46	.66	.17	.05
47	.68	.18	.10
48	.96	.30	-.38
49	.81	.11	.78
50	.67	.10	.09

Note. N = 1,483.

This means that, if scoring of the VAST were based on the loadings for the first factor, the test's reliability would be greater than .80.

When we extracted two factors for the respective subsets and rotated them to the varimax criterion, the second factors were not interpretable. Based on the eigenvalues and the two-factor solution, we tentatively concluded that there is only one substantial factor in the VAST. A visual inspection indicates that of the nine items that loaded at least .40 on the first factors (7, 9, 15, 24, 29, 40, 43, 44, and 48), eight are simple and concrete in design with no brush work. One item (24) is complex, containing abstract brush strokes.

Table 10

Descriptive Statistics for Three Subscales of the VAST

Subscale	No. items	Rel.	Average item-total corr.
Item-totals >.18[a]	21	.61	.22
Item-totals >.20[b]	16	.58	.23
Balance & harmony[c]	21	.56	.20

[a]Items 2, 3, 9, 15, 17, 6, 7, 23, 5, 11, 12, 24, 20, 28, 35, 38, 40, 44, 29, 43, and 48.

[b]Items 3, 9, 15, 7, 5, 11, 12, 24, 32, 35, 38, 40, 44, 29, 43, and 48.

[c]Items in which keyed designs clearly show superior balance and harmony are Items 1, 2, 3, 4, 5, 6, 7, 15, 18, 23, 25, 29, 32, 34, 35, 37, 38, 40, 43, 44, and 45.

The Rasch infit values for the VAST items ranged from -2.08 to 7.69. Three items exceeded the fit criterion of 3.0 (4, 16, and 21), and two of those are abstract brush stroked images. Item 21 is the only simple, concrete image that exceeded the fit criterion, and the p value (.58) indicates it is a difficult item. A comparison of item fit and item difficulty indicates that the best-fitting items are generally the easiest items on the test and tend to contain the simple concrete designs.

As we indicated above, the alpha reliability for this printed version of the VAST (.66) was substantially below the reliability of .84 that Götz et al. (1979) reported for photographic slide

projections with a group of college students. We speculate that the printed reproductions of the highly detailed brush-stroked items do not provide sufficient contrast of the fine detail associated with their balance and harmony for examinees to form reliable judgments, as opposed to the clear contrast between balance and harmony found in the simple and concrete items. Based on our results and the Spearman-Brown prophecy formula, a set of 60 printed VAST items consisting of the type we call simple and concrete would be expected to have a test reliability greater than .80.

Sex and age effects. The analysis of variance of VAST scores indicated a significant main effect for age, $F(6;1,365) = 5.14$, $p < .001$, and a significant interaction for sex and age, $F(6;1,365) = 2.95$, $p < .001$. The sex effect was not significant.

The largest difference between age groups was .41 standard deviation units, and the age effect accounted for a small proportion of the variance ($w^2 = .026$), equivalent to the effect of a correlation of .16. The age curve in Figure 10 indicates substantially increasing scores until the mid-twenties, followed by a plateau and then a modest decline through the succeeding adult years.

Because the sex by age interaction was significant (although w^2 was only .007), we also examined the age curves separately for males and females. The age curve for males was like the overall age curve but with larger effects, whereas females showed almost no age differences.

Summary.

1. Three of the items had p values equal to or greater than .95.

2. Item-total correlations were generally low. Only 13 items had item-totals over .20 (3, 5, 7, 9, 12, 15, 24, 29, 38, 40, 43, 44, and 48).

3. Alpha reliability was low, .66.

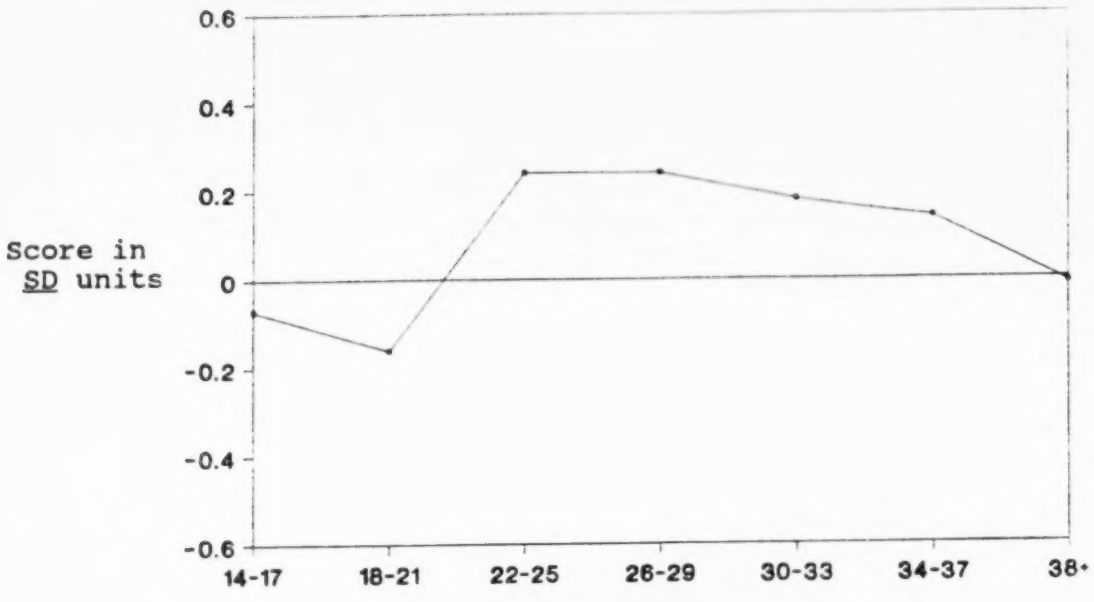
4. Two abstract items (4 and 16) and one concrete item (21) did not fit the Rasch measurement model.

5. In a comparison of item subsets, two groups of items with simple solid designs had average item-total correlations of .22 and .23, respectively.

6. A modest age effect accounted for a small proportion of the score variance. The age curve indicated increasing scores to the mid-twenties, then a plateau and decline through adulthood. No sex difference was found.

Figure 10

Age Curve for Visual Aesthetic Sensitivity Test



Note. Plotted values are not smoothed.

The Artistic Judgment Battery

Correlations Among the Artistic Judgment Tests

Pearson product-moment correlations among the four artistic judgment tests are displayed in Table 11. The between-test correlations ranged from -.28 to .30 with correlations of .19 between the two VDT scales and .61 between the PA scores. When corrected for attenuation, PA-C and PA-.67 correlated .80, suggesting little practical difference between the two scoring methods. The two largest between-test correlations both involved the DJT, .30 with VDT Simplicity and -.28 with PA-consensus. Thus, there is a modest degree of association between preferences for asymmetry (DJT) and Simplicity (VDT), both of which are expected to relate to overall artistic judgment. The tendency for DJT to be associated with preference for the less popular shapes when PA is scored for consensus is consistent with previous Foundation research and perhaps consistent with the earlier report that artists prefer the more-elongated to the golden-section proportions on PA.

As we expected, the VAST showed its strongest correlation (.19) with PA scored for consensus, suggesting that sensitivity to differences in balance and harmony on the VAST tends to associate with preferences for shapes that have broad appeal. This relationship is consistent with the desire of the test developers of the VAST to measure a general construct for "good taste."

One of the surprising findings is the negative correlation between VDT Uniformity and DJT. Although DJT and Uniformity both correlate positively with Simplicity, their between-test correlation is -.25.

Table 11

Correlations Among the Artistic Judgment Tests

Test[a]	DJT	SIM	UNI	PA-C	PA-.67	VAST
Design Judgment Test		30	-25	-28	-09	-08
Visual Designs Test, Simplicity			19	-14	NS	NS
Visual Designs Test, Uniformity				06	NS	08
Proportion Appraisal (Consensus scoring)					61	19
Proportion Appraisal (.67 scoring)						13
Visual Aesthetic Sensitivity Test						
Reliability	89	95	88	76	78	66

Note. Ns range from 1,446 to 1,650. Leading decimals omitted. Only correlations significant at $p < .05$ level displayed.

[a]Notation for tests: DJT (Design Judgment Test); SIM (Simplicity); UNI (Uniformity); PA-C (Proportion Appraisal: Consensus scoring); PA-.67 (Proportion Appraisal: .67 scoring); VAST (Visual Aesthetic Sensitivity Test).

Because the tests showed only minimal correlation with each other, we interpreted them as measuring distinct dimensions and did not perform a factor analysis.

Overall, the correlations in Table 11 indicate that the four tests measure to a large extent different aspects of visual preference. They do not converge on a single dimension that could be interpreted as quality of artistic judgment. It is possible that one of the tests represents general artistic judgment and the others do not (causing the minimal intertest correlations) or that there is a separate artistic judgment dimension to which the different tests studied here are only modestly related. Alternatively, artistic judgment may be a multidimensional phenomenon.

DISCUSSION

The overall goal of the artistic judgment project is to evaluate tests that measure aspects of visual preference that may be related to artistic judgment. The analyses reported here provide a foundation for that effort by elucidating the internal psychometric structures of the four tests under study. To the extent that these tests are reliable and unidimensional, they have the potential to be effective artistic judgment tests. Later reports in this series will address construct and occupational validation data for the tests.

The VDT was found to contain two major factors, and scales were formed for each of these factors. The two VDT scales and the 22-item version of the DJT showed good reliability. For both of its scoring procedures, PA showed reliabilities in the .70s. These coefficients are below the Foundation standard of .80 and are disappointing in view of the length of the test. The printed version of the VAST that we used here showed poor reliability (.66), although the slide version used in previous studies (Götz et al. 1979) had better reliability (.84). Also, as noted earlier, if the VAST were revised to contain only simple concrete designs, we estimate the reliability for 60 items would be over .80.

In terms of factor analysis, we found the DJT to contain only one factor. Thus, our study confirmed earlier research (Eysenck, 1967, 1972b) indicating high reliability and unidimensionality (or approximate unidimensionality) for this shortened version of the DJT. PA and the VAST each had one primary factor with minor secondary factors. As noted earlier, the VDT contained two primary factors, both of which we used to develop scales.

Although significant sex effects were obtained for several of the tests, only the DJT had a sex difference large enough to have implications for individual assessment. Females showed a

significantly stronger preference than males for asymmetrical designs. In this respect, females scored more like artists, based on previous research, than did males.

The age effects we found were small but fairly consistent. The widest range of means was on VDT Simplicity, .65 of a standard deviation. Each of the tests showed generally increasing scores through the teens and into the 20s followed by a leveling off resembling an adult plateau. The VDT, however, showed deviation from this pattern in that scores on both Simplicity and Uniformity increased from the early to the mid-30s to the 38+ group.

These age differences in adulthood were rather small and probably do not threaten the interpretation of these tests as measures of aptitudes. It is possible, however, that the post-25 trends on these tests reflect differences between birth cohorts rather than maturational effects--that is, persons born in the early 1950s may have greater preferences for simplicity and uniformity than persons born in the late 1950s. If so, then some of the tests may be affected in a limited way by culture and/or temporal fashions. Longitudinal research would enable us to distinguish between cohort and maturational effects.

One aspect of the results that relates directly to our primary research goal, measuring artistic judgment aptitude, is the intertest correlations. Consistent with prior research, these tests show only low correlations with each other, implying that they measure separate dimensions of visual preference. An examination of the direction of the correlations, however, reveals a pattern that is theoretically meaningful. First, the VAST, one of the tests for which validation research with artists has previously been conducted and which is believed to measure Eysenck's T factor for artistic judgment, showed its highest positive correlation with Proportion Appraisal-consensus scoring, a score that reflects agreement with a general consensus regarding the attractiveness of shapes. This finding provides some evidence in support of a general aptitude for artistic judgment and is consistent with O'Connor's perspective regarding a global capacity for artistic sensitivity.

Second, Simplicity, the first factor from the VDT, which is conceptually linked to Eysenck's K factor, showed its highest positive correlation with the DJT. This association has interesting implications because the DJT has had extensive validation linking it to artistic preference. This relationship implies that our attempt to construct a new scale (the VDT) on which artists and nonartists differ may have been successful.

A third, unexpected finding in our analysis of intertest correlations is the negative association between VDT Uniformity and the DJT. Uniformity is positively associated with VDT Simplicity, and we expected a somewhat similar relationship

between Uniformity and DJT. What we found is an inverse relationship. If we assume that the DJT has validity as a test of artistic preference, the association between Uniformity and DJT suggests that we may have discovered another factor on which artists differ from nonartists (in addition to their preference for less-complex over more-complex designs on Simplicity), which, not surprisingly, could be a preference for less-uniform designs. When presented with a choice between relatively ordered, uniform, simple patterns and more-chaotic, more-complex ones, the association between DJT and Uniformity suggests that artists tend to choose the more-complex patterns.

In our next report in this series, we will address these issues again after examining the correlations of these tests with the Foundation's standard aptitude battery, occupational interest scales, and an artistic activities questionnaire. In a later report, we will conclude this series on artistic judgment with an analysis of the visual preferences of professional artists.

GENERAL SUMMARY AND CONCLUSIONS

1. The highest reliability was obtained for the Simplicity scale on the VDT (r_{tt} = .95). The lowest reliability was for the VAST (.66), which was used in its printed version. The reliabilities for DJT, VDT Uniformity, and Consensus and .67 for Proportion Appraisal were .89, .88, .76, and .78, respectively. Relative to use in the standard Foundation battery, the DJT and the two VDT scales have adequate (in fact, good) reliability. PA would need to be improved, and the VAST has poor reliability.

2. A one-factor solution (with residual covariation for PA and the VAST) was adequate for all the tests except the VDT, which had two factors.

3. All the tests approximate the quantitative requirements for a linear measurement model (i.e., the Rasch model).

4. On Proportion Appraisal, the shape chosen as the best most frequently had a proportion of .75. The shape chosen most frequently as the worst was based on .50.

5. Only the DJT showed a substantial sex difference. On that test, females scored .43 of a standard deviation higher in the direction of preference for asymmetry.

6. In terms of age, the effects were generally small, amounting to a few tenths of a standard deviation. Scores tended to increase into adulthood and then level off. VDT Simplicity and Uniformity scores increased again for the 38+ age group, indicating a greater preference for simple and uniform designs.

7. The intertest correlations are modest, suggesting the tests measure relatively unique aspects of visual preference.

8. The intertest correlations show patterns consistent with previous factor analytic research.

REFERENCES

Anderson, R. G. (1951). A note on the McAdory and Meier Art Tests in counseling. Educational and Psychological Measurement, 11, 81-86.

Andreasen, N. C. (1987). Creativity and mental illness: Prevalence rate in writers and their first-degree relatives. American Journal of Psychiatry, 144, 1288-1292.

Attneave, F. (1954). Informational aspects of visual perception. Psychological Review, 61, 183-193.

Attneave, F. (1956). The quantitative study of shape and pattern perception. Psychological Bulletin, 6, 452-471.

Attneave, F. (1957). Physical determinants of the judged complexity of shapes. Journal of Experimental Psychology, 53, 221-226.

Attneave, F. (1959). Stochastic composition processes. Journal of Aesthetics and Art Criticism, 17, 503-510.

Barrett, H. (1949). An examination of certain standardized art tests to determine their relation to classroom achievement. Journal of Educational Research, 42, 398-400.

Barron, F. (1953). Complexity-simplicity as a personality dimension. Journal of Abnormal and Social Psychology, 48, 163-172.

Barron, F. (1965). The psychology of creativity. In F. Barron et al. (Eds.), New directions in psychology II (pp. 3-134). New York: Holt, Rinehart & Winston.

Barron, F., & Welsh, G. (1952). Artistic perception as a possible factor in personality style: Its measurement by a figure preference test. Journal of Psychology, 33, 199-203.

Berlyne, D. E. (1971). Aesthetics and psychobiology. New York: Appleton-Century-Crofts.

Berlyne, D. E. (1974). Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation. Washington, DC: Hemisphere.

Biaggio, M. K., & Supplee, K. A. (1983). Dimensions of aesthetic perception. Journal of Psychology, 114, 29-35.

Birkhoff, G. P. (1932). Aesthetic measure. Cambridge, MA: Harvard University Press.

Birkhoff, G. P. (1956). Mathematics of aesthetics. In J. Newman (Ed.), The world of mathematics (Vol. 4, pp. 2185-2195). New York: Simon & Schuster.

Bolton, E. B. (1955). Brief evaluation of two tests of aesthetic judgment. Peabody Journal of Education, 32, 211-223.

Bosanquet, B. (1917). A history of aesthetics. London: G. Allen & Unwin.

Brighouse, G. (1939a). A study of aesthetic appreciation. Psychological Monographs, 51, 1-22.

Brighouse, G. (1939b). Variability in preferences for simple forms. Psychological Monographs, 51, 68-74.

Bulley, M. H. (1951). Art and everyman (Vol. I). London: Batsford.

Carroll, H. A. (1933). What do the Meier-Seashore and the McAdory Art Tests measure? Journal of Educational Research, 26, 661-665.

Carroll, H. A., & Eurich, A. C. (1932). Abstract intelligence and art appreciation. Journal of Educational Psychology, 23, 214-220.

Champa, K. S. (1985). Mondrian studies. Chicago: University of Chicago Press.

Chan, J., Eysenck, H. J., & Götz, K. O. (1980). A new visual aesthetic sensitivity test: III. Cross-cultural comparisons between Hong Kong children and adults, and English and Japanese samples. Perceptual and Motor Skills, 50, 1325-1326.

Child, I. L. (1964). Observations on the meaning of some measures of esthetic sensitivity. Journal of Psychology, 57, 49-64.

Child, I. L. (1965). Personality correlates of esthetic judgment in college students. Journal of Personality, 33, 476-511.

Coffman, W. E. (1957). The Tests of Developed Abilities: Development of the tests. College Board Review, 31, 6-10.

Cohen, J., & Cohen, P. (1975). Applied multiple regression/ correlation analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Conoley, J. C., & Kramer, J. J. (1989). (Eds.), The Tenth Mental Measurement Yearbook. Lincoln, Nebraska: Buros Institute of Mental Measurements

Crannell, C. W. (1953). The validity of certain measures of art appreciation in relation to a drawing task. Journal of Psychology, 35, 131-142.

Crawford, A. B., & Burnham, P. S. (1946). Forecasting college achievement. New Haven: Yale University Press.

Csikszentmihalyi, M., & Getzels, J. (1973). The personality of young artists: An empirical and theoretical exploration. British Journal of Psychology, 64, 91-104.

Dewar, H. (1938). A comparison of tests of artistic appreciation. British Journal of Educational Psychology, 8, 29-49.

Dorfman, D. D. (1965). Esthetic preference as a function of pattern information. Psychonomic Science, 3, 85-86.

Dyer, H. S. (1957). The Tests of Developed Abilities: The tests' background. College Board Review, 31, 5-6.

Eisenman, R. (1966). The association of random shapes revisited. Psychonomic Science, 6, 397-398.

Eisenman, R., & Rappaport, J. (1967). Complexity preference and semantic differential ratings of complexity-simplicity and symmetry-asymmetry. Psychonomic Science, 7, 147-148.

Ellis, H. A. (1926). A study of British genius. London: Houghton-Mifflin.

Eysenck, H. J. (1940a). The general factor in aesthetic judgements. British Journal of Psychology, 31, 94-102.

Eysenck, H. J. (1940b). Some factors in the appreciation of poetry and their relation to tempermental qualities. Character & Personality, 9, 160-167.

Eysenck, H. J. (1941a). Type factors in aesthetic judgements. British Journal of Psychology, 31, 262-270.

Eysenck, H. J. (1941b). The empirical determination of an aesthetic formula. Psychological Review, 48, 83-92.

Eysenck, H. J. (1942). The experimental study of the "good gestalt"--a new approach. Psychological Review, 49, 344-364.

Eysenck, H. J. (1957). Sense and nonsense in psychology. Baltimore: Penguin.

Eysenck, H. J. (1967). Factor analytic study of the Maitland Graves Design Judgement Test. Perceptual and Motor Skills, 24, 73-74.

Eysenck, H. J. (1968). An experimental study of aesthetic preference for polygonal figures. Journal of General Psychology, 79, 3-17.

Eysenck, H. J. (1970). An application of the Maitland Graves Design Judgement Test to professional artists. Perceptual and Motor Skills, 30, 598-590.

Eysenck, H. J. (1971). Factors determining aesthetic preferences for geometrical designs and devices. British Journal of Aesthetics, 11, 154-166.

Eysenck, H. J. (1972a). Personal preferences, aesthetic sensitivity and personality in trained and untrained subjects. Journal of Personality, 40, 544-557.

Eysenck, H. J. (1972b). The development of aesthetic sensitivity in children. Journal of Child Psychology and Psychiatry, 13, 1-10.

Eysenck, H. J. (1981). Aesthetic preferences and individual differences. In D. O'Hare (Ed.), Psychology and the Arts (pp. 76-101). Brighton, Sussex, United Kingdom: Harvest Press.

Eysenck, H. J. (1983). A new measure of 'good taste' in visual art. Leonardo, 16, 229-231.

Eysenck, H. J. (1988). Personality and scientific aesthetics. In F. H. Harley and R. W. Neperud, (Eds.), The Foundations of Aesthetics, Art, & Art Education (pp. 117-160). New York: Praeger.

Eysenck, H. J., & Castle, M. (1970). Training in art as a factor in the determination of preference judgements for polygons. British Journal of Psychology, 61, 65-81.

Eysenck, H. J., & Castle, M. (1971). A comparative study of artists and nonartists on the Maitland Graves Design Judgment Test. Journal of Applied Psychology, 55, 389-392.

Eysenck, H. J., & Eysenck, S. B. G. (1964). Eysenck Personality Inventory. London: University of London Press.

Eysenck, H. J., & Eysenck, S. B. G. (1976). Psychoticism as a dimension of personality. London: Hodder & Stoughton.

Eysenck, H. J., & Turnstall, O. (1968). La personalité et l'esthetique des formes simples. Sciences de L'Art, 5, 3-9.

Fechner, G. T. (1865). Weber die Fragen des golden Schnitts. Archiv. f. d. Zeichenenden Kuenste (Nauman-Weigel), 11, 100-112.

Fechner, G. T. (1876). Vorschule der aesthetik. Leipzig: Breitkopf & Hartel.

Francés, R., & Voillaume, H. (1964). Une composante du judement pictural: la fidélité de la représentation. Psychologie Francaise, 9, 241-256.

Fritzky, F. J. (1963). Aesthetic preference for abstract designs as a function of their perceived complexity. Educational Testing Service Bulletin, 27, 1-41.

Garner, W. R. (1970). Good patterns have few alternatives. American Scientist, 58, 34-42.

Getzels, J. (1976). The Creative Vision: The evaluation of works of art in the creative vision. New York: John Wiley.

Götz, K. O. (1981). VAST: Visual Aesthetic Sensitivity Test. Düsseldorf: Concept Verlag.

Götz, K. O., & Götz, K. (1974). The Maitland Graves Design Judgment Test judged by 22 experts. Perceptual and Motor Skills, 39, 261-262.

Götz, K. O., & Götz, K. (1979a). Personality characteristics of professional artists. Perceptual and Motor Skills, 49, 327-334.

Götz, K. O., & Götz, K. (1979b). Personality characteristics of successful artists. Perceptual and Motor Skills, 49, 919-924.

Götz, K. O., Lynn, R., Borisy, A. R., & Eysenck, H. J. (1979). A new visual sensitivity test: I. Construction and psychometric properties. Perceptual and Motor Skills, 49, 795-802.

Goude, G. (1972). A multidimensional scaling approach to the perception of art II. Scandinavian Journal of Psychology, 13, 258-271.

Granger, G. W. (1955). An experimental study of colour preferences. Journal of General Psychology, 52, 21-35.

Graves, M. (1948). Design Judgment Test: 1948 manual. New York: The Psychological Corporation.

Haber, R. N., & Hershenson, M. (1965). Effects of repeated brief exposures on the growth of a percept. Journal of Experimental Psychology, 69, 40-46.

Harsh, C. M., Beebe-Center, J. G., & Beebe-Center, R. (1939). Further evidence regarding preferential judgment of polygon forms. Journal of Psychology, 7, 342-350.

Hill, N. (1968). Farnsworth-Munsell 100-Hue Test. Paper presented at the research meeting of the Johnson O'Connor Research Foundation.

Hochberg, J., & McAlister, E. (1953). A quantitative approach to figural "goodness". Journal of Experimental Psychology, 46, 361-364.

Iwawaki, S., Eysenck, H. J., & Götz, K. O. (1979). A new visual sensitivity test (VAST): II. Cross cultural comparisons between England and Japan. Perceptual and Motor Skills, 49, 859-862.

Juda, A. (1949). The relationship between higher mental capacity and psychic abnormalities. American Journal of Psychiatry, 106, 296-307.

Kellet, K. R. (1939). A gestalt study of the functions of unity in aesthetic perception. Psychological Monographs, 51, 23-51.

Kline, L. W., & Carey, G. L. (1922). The Kline-Carey measuring scale for free-hand drawing. Baltimore: Johns Hopkins Press.

Koffka, K. (1935). Principles of gestalt psychology. New York: Harcourt & Brace.

Köhler, W. (1920). Die physischen Gestalten in Ruhe und im stationaren zustand. Erlangen: Weltkreisverlag.

Külpe, O. (1903). Beitrag zur experimentaellen aesthetik. American Journal of Psychology, 14, 215-231.

Lemoine, S. (1987). Mondrian and De Stijl. (Translated by Charles Lynn Clark), New York: Universe Books.

Likert, R., & Quasha, W. H. (1948). The Revised Minnesota Paper Form Board Test, manual. New York: The Psychological Corporation.

Lukens, H. T. (1896). A study of children's drawing. Pedagogical Seminary. 4.

Manuel, H. T. (1919). Talent in drawing. School and Home Education Monograph, 3, Bloomington, IL: Public School Publishing Company.

McAdory, M. (1929). The construction and validation of an art test. New York: Bureau of Publications, Columbia University.

McNeil, T. F. (1971). Prebirth and postbirth influence on the relationship between creative ability and mental illness. Journal of Personality, 39, 391-406.

Meier, N. C. (1928). A measure of art talent. Psychological Monographs, 39, 184-199.

Meier, N. C. (1942a). Art in human affairs. New York: McGraw-Hill.

Meier, N. C. (1942b). The Meier Art Tests: Part I. Art judgment. Iowa City: State University of Iowa.

Meier, N. C. (1942c). Meier Art Judgment Test. New York: The Psychological Corporation.

Meier, N. C., & Seashore, C. E. (1930). The Meier-Seashore Art Judgment Test. Iowa City: Bureau of Educational Research and Service, University of Iowa.

Millman, M., & Chang, T. (1966). Intercorrelations among three widely used art tests. Perceptual and Motor Skills, 23, 1002.

Munsinger, H., & Kessen, W. (1964). Uncertainty, structure, and preference. Psychological Monographs, 78, 1-24.

Neale, J. M., & Cromwell, R. L. (1969). Preference for complexity in acute schizophrenics. Journal of Consulting and Clinical Psychology, 33, 245-246.

Noll, A. M. (1966). Human or machine: A subjective comparison of Piet Mondrian's "Composition With Lines" (1917) and a computer generated picture. The Psychological Record, 16, 1-10.

Noll, A. M. (1972). The effects of artistic training on aesthetic preferences for pseudo-random computer-generated patterns. The Psychological Record, 22, 449-464.

Norusis, M. J. (1988a). SPSS/PC+ V2.0 base manual. Chicago: SPSS.

Norusis, M. J. (1988b). SPSS/PC+ Advanced Statistics V2.0. Chicago: SPSS.

Nunnally, J. C. (1967). Psychometric theory. New York: McGraw-Hill.

O'Connor, J. (1940). Unsolved business problems. New York: Johnson O'Connor Research Foundation.

O'Connor, J. (1948). The unique individual. New York: Johnson O'Connor Research Foundation.

Osborne, J. W., & Farley, F. W. (1970). The relationship between aesthetic performance and visual complexity in abstract art. Psychonomic Science, 19, 69-70.

Peel, E. A. (1944). On identifying aesthetic types. British Journal of Psychology, 35, 61-69.

Pickford, R. W. (1948). 'Aesthetic' and 'technical' factors in artistic appreciation. British Journal of Psychology, 38, 135-141.

Research Memorandum 1989-1. (1989). Item construction codes and Rasch rating scale calibrations for 45 visual designs. N. Bezruczko. Chicago: Johnson O'Connor Research Foundation.

Research Memorandum 1989-2. (1989). Rasch item calibrations for the artistic judgment battery. N. Bezruczko. Chicago: Johnson O'Connor Research Foundation.

Research Memorandum 1989-3. (1989). Item pairs for 84 VDT forced-choice items. N. Bezruczko. Chicago: Johnson O'Connor Research Foundation.

Rosenblatt, E., & Winner, E. (1988). Is superior visual memory a component of superior drawing ability? In L. Obler and D. Fein (Eds.), The exceptional brain. New York: Guilford.

Schroth, M. L. (1967). Spatial aptitude and its relationship to art judgment. Perceptual and Motor Skills, 24, 746.

Siceloff, M. M, & Woodyard, E. (1933). McAdory Art Test (revised). New York: Bureau of Publications, Teachers College, Columbia University.

Statistical Bulletin 1975-6. (1975). Design Judgment research, Worksample 654 AA leading to Forms B and C: Verification of factor analysis by H. J. Eysenck. C. Quimby & B. F. Leshikar. Boston: Johnson O'Connor Research Foundation.

Statistical Bulletin 1980-5. (1980). Orientation of Wiggly Block Worksample 6 related to degree of difficulty. Chicago: Johnson O'Connor Research Foundation.

Statistical Bulletin 1986-5. (1986). Research proposal: Instrument for artistic judgment. N. Bezruczko. Chicago: Johnson O'Connor Research Foundation.

Taylor, R. E., & Eisenman, R. (1964). Perception and production of complexity by creative art students. Journal of Psychology, 57, 239-242.

Technical Report 23. (1939). Factors in art appreciation test. Worksample 172. S. P. Horton. New York: Johnson O'Connor Research Foundation.

Technical Report 44. (1939). Two experiments in the simplification of the measure of art appraisal. S. P. Horton. New York: Johnson O'Connor Research Foundation.

Technical Report 99. (1942). A further experiment in the measurement of Proportion Appraisal. S. J. Lindstrum. New York: Johnson O'Connor Research Foundation.

Technical Report 251. (1946). A study of Proportion Appraisal Worksample 235 Form B. L. J. Quesnel. New York: Johnson O'Connor Research Foundation.

Technical Report 792. (1973). Design Judgment research, Worksample 654 Form A, leading to Form AA: First purification. I. C. Shambaugh. Chicago: Johnson O'Connor Research Foundation.

Technical Report 1988-1. (1988). Artistic judgment: A review of research. N. Bezruczko. Chicago: Johnson O'Connor Research Foundation.

Terman, L., & Oden, M. H. (1947). The gifted child grows up. In Genetic studies of genius (Vol. 4). Stanford, CA: Stanford University Press.

Test Information Bulletin 1980-5. (1980). New form letters; New Worksample Numbers, Worksample 691 A. Chicago: Johnson O'Connor Research Foundation.

Test Information Bulletin 1986-10. (1986). Rating Scale of Visual Designs: Worksample 735 A*. Chicago: Johnson O'Connor Research Foundation.

Thorndike, E. L. (1913). The measurement of achievement in drawing. Teachers College. New York: Columbia University.

Thorndike, E. L. (1916). Tests of esthetic appreciation. Journal of Educational Psychology, 7, 509-522.

Thurstone, L. L. (1927a). Equally often noticed differences. Journal of Educational Psychology, 18, 289-293.

Thurstone, L. L. (1927b). Mental unit of measurement. Psychological Review, 34, 415-423.

Thurstone, L. L. (1927c). Psychophysical analysis. American Journal of Psychology, 38, 368-389.

Thurstone, L. L. (1931). The measurement of social attitudes. Journal of Abnormal and Social Psychology, 26, 249-269.

Thurstone, L. L. (1948). Psychological implications of factor analysis. American Psychologist, 3, 402-408.

Thurstone, L. L. (1954). The measurement of values. Psychological Review, 61, 47-58.

Trembly, D. (1974). Learning to use your aptitudes. San Luis Obispo, CA: Erin Hills Publishers.

Vanderplas, J. M., & Garvin, E. A. (1959). The association value of random shapes. Journal of Experimental Psychology, 57, 147-154.

Vitz, P. C. (1966). Preference for different amounts of visual complexity. Behavioral Science, 11, 105-114.

Welsh, G. S. (1949). A projective figure-preference test for diagnosing psychopathology: I. Preliminary investigation. Unpublished doctoral thesis, University of Minnesota.

Welsh, G. S. (1975). Creativity and Intelligence: A personality approach. Chapel Hill, NC: Institute for Research in Social Science, University of North Carolina.

Welsh, G. S., & Barron, F. (1963). Barron-Welsh Art Scale. Palo Alto, CA: Consulting Psychologists.

Winner, E. (1988). Visual and spatial abilities in young artists. Paper presented at the Symposium on Art and the Brain, May 12-14, Chicago.

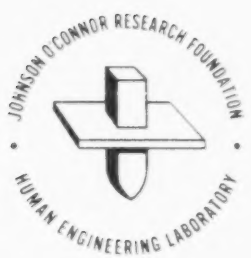
Woods, W. A. (1953). Visual-form-space manipulation and its relation to art participation. Dissertation Abstracts, 13, 129.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.

Wright, B. D., & Schulz, E. M. (1988). BANK: A Rasch program for partial credit scoring. Chicago: University of Chicago, Department of Education, Statistical Laboratory.

Wright, B. D., & Stone, M. (1979). Best test design. Chicago: MESA Press.

Wright, B. D., Congdon, R. T., & Rossner, M. (1987). MSCALE: A Rasch program for rating scale analysis. Chicago: University of Chicago, Department of Education, Statistical Laboratory.


JOHNSON O'CONNOR RESEARCH FOUNDATION
HUMAN ENGINEERING LABORATORY